Application of Information Theory and Learning to Network and Biological Tomography

A Thesis
Presented to
The Academic Faculty

by

Rajesh Narasimha

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

School of Electrical and Computer Engineering
Georgia Institute of Technology
December 2007

Application of Information Theory and Learning to Network and Biological Tomography

Approved by:

Dr. Steven McLaughlin, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Chuanyi Ji, Co-Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Aaron Lanterman
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Bonnie Heck Ferri
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Alexander Gray
College of Computing
Georgia Institute of Technology

Dr. Sriram Subramaniam
National Cancer Institute
National Institutes of Health

Date Approved: 8 November 2007

To Doddappa, Doddamma and Rashmy

ACKNOWLEDGEMENTS

During the final days of my MS studies, I would persistently meet with my advisor, Dr. Raghuveer Rao about my career goals. He would listen patiently to my thoughts, and always provide ever-insightful advice. I would say he is partly responsible for my decision to pursuing my Ph.D. degree. I sincerely thank him for being a "father figure" and for his immense patience. Dr. Steven McLaughlin was the other person who motivated me to join Georgia Tech and helped me thread my way through my Ph.D. years in a supportive way. It was a real pleasure working with him and I thank him for being an excellent advisor and a great person. These two people have made a lot of difference in my life. I would like to thank Dr. Chuanyi Ji for her support both technically and financially. I also would like to thank Dr. Aaron Lanterman, Dr. Bonnie Heck Ferri, Dr. Sriram Subramaniam, and Dr.Alexander Gray for serving on my defense committee. Dr. Bonnie Heck Ferri was very helpful as a ECE graduate chair. Special thanks to Aaron Lanterman for proof reading my thesis. Many thanks are due to Sriram Subramaniam of NCI, NIH for being extremely supportive and introducing me to research problems in biology. It was a real pleasure to work in his lab, with great people, as a pre-doctoral fellow.

Life has taken a lot of twists and turns and it would have been much worse if it had not been for three people, Doddamma, Doddappa, and Rashmy. They have been very encouraging, loving, and caring throughout my long journey, and have set an example for kindness and simplicity that I would like to follow in the years to come. Rashmy, came into my life when I was in a low throughput phase, and it was a mere "fairy tale" I guess, although it took two weekends, some books, 2000 miles and 10 gallons of fuel to tie the knot. She has been full of life ever since I met her,

and I thank her for being there through all the ups and downs of life. Of course, my parents and sisters always kept asking me the same question: "When will you finish studying?" They are partly responsible for me graduating, and I thank them for their support and patience.

Finishing a Ph.D is a marathon, they say, and I have been helped and supported along the way by a great number of people. First, I would like to thank my lab mates Guanglei, Zesheng, Noo, and Sung-eok. It was a pleasure working with them. Thanks are also due to the staff at ECE, Georgia Tech, notably Cordai Farrar, Lisa Gardner, Suzzette Willingham, Marilou Mycko, and Robert James at the co-op office, who were very helpful in the administrative details of the PhD process, travelling to conferences, etc. I would also like to thank Martin Kessel, Ellen Hillburn and Karen Knight of NCI, NIH for all their help regarding the pre-doctoral fellowship appointment. I thank all my collaborators, especially Dr. Guillermo Sapiro and his student Iman Aganj from UMN, Mario, Adam, Daniel, Kim from NCI, Alexander Gray from CoC (for our discussions over coffee at Starbucks), and Matthieu Bloch for our discussions.

My stay at Tech was truly enjoyable, and thanks are due to many of my friends in Atlanta and elsewhere, including the GCATT gang, Aravind and Badri, who usually hunt in pairs, for entertainment at the cost of others, with their usual scapegoats Souvik, Shayan, Aru and me. I also thank Karthik, Sridhar, Pradnya, and Srini for putting up with me and my research. Our house, "1186 Holly", the place for the hungry, the homeless, and the distressed; I need to thank Badri for finding this heavenly place. I think, at some point in time, all of the above gang members have stayed there. Thanks also must go to Yogi, Para, Nupur, Deepak, Ponnappan and Vellambi for all those badminton games and dinner outings.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Studying the internal characteristics of a network using measurements obtained from end-hosts is known as network tomography. The foremost challenge in measurement-based approaches is the large size of a network, where only a subset of measurements can be obtained because of the inaccessibility of the entire network. As the network becomes larger, a question arises as to how rapidly the monitoring resources (number of measurements or number of samples) must grow to obtain a desired monitoring accuracy. Our work studies the scalability of the measurements with respect to the size of the network. We investigate the issues of scalability and performance evaluation in IP networks, specifically focusing on fault and congestion diagnosis. We formulate network monitoring as a machine learning problem using probabilistic graphical models that infer network states using path-based measurements. We consider the theoretical and practical management resources needed to reliably diagnose congested/faulty network elements and provide fundamental limits on the relationships between the number of probe packets, the size of the network, and the ability to accurately diagnose such network elements. We derive lower bounds on the average number of probes per edge using the variational inference technique proposed in the context of graphical models under noisy probe measurements, and then propose an entropy lower (EL) bound by drawing similarities between the coding problem over a binary symmetric channel and the diagnosis problem. Our investigation is supported by simulation results. For the congestion diagnosis case, we propose a solution based on decoding linear error control codes on a binary symmetric channel for various probing experiments. To identify the congested nodes, we construct a graphical model, and infer congestion using the belief propagation algorithm.

In the second part of the work, we focus on the development of methods to automatically analyze the information contained in electron tomograms, which is a major challenge since tomograms are extremely noisy. Advances in automated data acquisition in electron tomography have led to an explosion in the amount of data that can be obtained about the spatial architecture of a variety of biologically and medically relevant objects with sizes in the range of 10-1000 nm.

A fundamental step in the statistical inference of large amounts of data is to segment relevant 3D features in cellular tomograms. Procedures for segmentation must work robustly and rapidly in spite of the low signal-to-noise ratios inherent in biological electron microscopy. This work evaluates various denoising techniques and then extracts relevant features of biological interest in tomograms of *HIV-1* in infected human macrophages and *Bdellovibrio* bacterial tomograms recorded at room and cryogenic temperatures. Our approach represents an important step in automating the efficient extraction of useful information from large datasets in biological tomography and in speeding up the process of reducing gigabyte-sized tomograms to relevant byte-sized data. Next, we investigate automatic techniques for segmentation and quantitative analysis of mitochondria in *MNT-1 cells* imaged using ion-abrasion scanning electron microscope, and tomograms of *Liposomal Doxorubicin* formulations (*Doxil*), an anticancer nano-drug, imaged at cryogenic temperatures. A machine learning approach is formulated that exploits texture features, and joint image block-wise classification and segmentation is performed by histogram matching using a *nearest neighbor* classifier and chi-squared statistic as a distance measure.

CHAPTER 1

INTRODUCTION

Measurement-based network monitoring has been widely used for diagnosis of network failures and operational performance degradation [1-4]. One of the less-studied key challenges of measurement-based network monitoring is the scalability issue, i.e., how to accurately estimate the underlying network states using the limited amount of available network monitoring resources (the number of probe-packet measurements). The problem of determining network status, e.g., connectivity, loss, congestion, and delay, at either a node or link level, is challenging in such probing scenarios because of large network sizes.

The first part of this thesis investigates the following scalability issues: (a) What is the number of measurements needed to determine the network states? and, (b) How rapidly should the monitoring resources required grow with the size of a network? For instance, suppose m measurements are sufficient to obtain the accurate estimates of network states for a small network with n nodes. If this network size becomes κn, where κ is a positive integer, would the number of measurements required become κm, $(\kappa n)^s; s \geq 2$ or $(2n)^\kappa$? Scalability is the growth rate of the number of measurements (i.e., probe packets) with respect to the network size that are required to monitor the network states given a performance bound [5]. Here, performance refers to the accuracy of network monitoring. Determining a trade-off between scalability and performance is a challenge. In addition to the growth rate, the actual constant that multiplies the growth function is of practical importance; together with the growth function, it provides an explicit number of measurements needed to infer the state of the network. Machine learning theory provides a framework based on probabilistic graphical models to derive theoretical limits and to statistically infer the network states using computationally efficient algorithms. Thus, the approaches used in this paper address the inference problem across the boundaries of machine learning and networking.

To motivate this work, we provide here some practical scenarios where scalability plays an important role. The Skitter tool from caida.org [6] gathers connectivity information, round trip time (RTT), and path data from source monitors scattered throughout the United States. To discover 131476 nodes and 279799 edges, Skitter uses approximately 54 million probes (193 probes per edge) [7]. Ideally, Skitter requires that the packet overhead on the existing traffic and the number of packets sent be minimized. Another example is the Resilient Overlay Network (RON) [8] architecture, which detects and recovers from path outages. RON suffers from scalability as the traffic growth is of $O(\Lambda^2)$, where Λ represents the nodes on the overlay network. Probes are intrusive and they impose cost in terms of additional network and computational load because the probe measurements have to be collected and analyzed. Hence, an ideal goal is to be able to reliably localize network problems in a large network with little overhead, i.e., the number of probes added to the traffic. Prior research in measurement-based network monitoring has been mostly focused on developing algorithms on "how to monitor" Our work focuses on two aspects: (a) fundamental limits on scalability and performance issues that ask, what can be done?, using a finite number of measurements, and (b) a monitoring approach that uses a scalable number of measurements. These two aspects have been less investigated in previous work on measurement-based network monitoring.

We study scalability and performance by focusing on two aspects of network monitoring namely failure and congestion diagnosis. Measurement-based diagnosis includes detection and localization in the interior of a network using path-based end-end probe measurements, where probe packets are sent from sources to destinations through network routes (path). In our approach, we assume that source routing is supported, which essentially means that we can specify the probing paths in advance. Simple examples of probes to analyze the status of a path are traceroute/ping and bandwidth monitoring tools such as the pipechar [9]. Researchers have investigated both passive and active probing methods to detect and mitigate network problems. For example, passive measurements can be obtained using network tools that are built into routers such as SNMP (Simple Network Management Protocol) [10], Netflow [11] and RMON [12] to monitor the internal status of the network. Typically, routers have to be polled to collect network statistics such as delay, loss, and available bandwidth. To obtain useful monitoring

information, a large amount of data needs to be collected, which can be prohibitive if the network size is large. For instance, studies have shown that periodic polling of Cisco-4000 series Netflow-enabled routers on a local network decreased the throughput by as much as 15-20% [1]. In an inter-domain setting, the domains are managed by multiple ISPs and the service providers do not disclose confidential information about their domains. Hence, active probing may be the only convenient choice when the interior of a network is not assessable directly [13]. Detection and localization may also be made easier using end-end measurements, since they contain global information on a network.

In machine learning theory, the number of measurements/samples needed to obtain a desired performance is known as sample complexity, which has been well studied [14]. However, the theory is not directly applicable to path-based end-end measurements, because path-based measurements are spatially and statistically dependent. For example, a failed link can cause end-end packet losses at multiple paths that traverse the link. Such a spatial dependence is necessary for failure and congestion diagnosis. Prior related works (see section 2.1) are focused on specific types of networks or conditions, and they motivate this work to provide a general approach for scalability of path-based diagnosis for IP networks. We formulate the detection and localization problem using graphical models representing the spatially dependent path-based measurements for analytical tractability and computational efficiency. Although many inference algorithms have been previously developed for path-based measurements [5, 13, 15-19], there have only been a few studies quantifying the amount of management resources needed [5, 20]. Hence, we study real-time inference algorithms that are robust, scalable, and have low computational complexity.

Our investigations are summarized as follows.

(a) We provide a general scalability definition and then study fundamental limits on the relationship between the number of probe packets, size of the network, and the ability to diagnose either failed links or congested network elements. Using graphical models that characterize the spatial and statistical dependence in path-based probe measurements, we investigate an approximate inference algorithm that is computationally efficient.

(b) Variational inference [21, 22] and source-coding techniques are applied to failure diagnosis to derive a lower bound on the average number of observations (m_f) required in the presence of

noise in the probe outcomes given the network size n_f. In the network diagnosis problem, subscripts "*f* "and "*c*" denote failure and congestion, respectively. The "noise" in our approach refers to the error (binary inversion) in the probe measurement at the end hosts. We show that if the number of observations is less than m_f, we are guaranteed not to achieve a zero diagnosis error asymptotically, where the performance measure is quantified in terms of the most probable explanation (MPE) error and the bit-error rate (BER). We further simplify the lower bound by considering a special case that imposes an additional constraint that all probe route lengths are equal, and show that the average number of measurements can be lower bounded as $m_f \geq c_f n_f$, where $c_f = \log(1/\phi_f)/c_{f_1}$, $\phi_f = \max(\rho_f, 1-\rho_f)$, ρ_f is the probability of faults, and c_{f_1} is a scaling constant, which is explicitly characterized in both our approaches. Thus, the lower bound provides a linear growth rate with respect to the size of the network and an expression for the constant c_f. Then, we examine the growth rate of the average number of measurements as the noise parameters increase, investigate scalable scenarios, and obtain non-scalable cases. The scalability results are verified using simulations. In the end, we provide insights on optimum probe route length both in the presence and absence of noise.

(c) By mapping the noisy network probing to a binary symmetric channel (BSC) in source coding, we develop a lower bound on m_f in terms of the entropy of ρ_f, and the noise parameters (ρ_1 and ρ_2), as $m_f / n_f \geq H(\rho_f)/H(\rho_{out})$, where $\rho_{out} = (1-\bar{\rho})\rho_1 + \bar{\rho}(1-\rho_2)$, $\bar{\rho} = (1-\rho_f)^r$, and *r* is the route length. We then compare the lower bound obtained using variational methods against the entropy lower (EL) bound, and show that the bound is tight for a certain range of noise values.

Next, we consider the congestion localization problem as a minor form of failure, which is dealt with in a two-fold fashion.

(d) Initially, the number of observations, denoted as m_c, is fixed, and we evaluate the maximum fraction (ρ_c) of the congested nodes that can be localized without error. Secondly, we examine the growth rate of the number of measurements as the percentage of congested nodes increases. This is accomplished by deriving theoretical bounds on the growth rate of m_c by relating the

congestion localization in networks using a bit-flipping mechanism to the problem of decoding linear error correcting codes (ECC) over a BSC channel in coding theory. In particular, we show that the number of measurements $m_c \geq c_c n_c'$, where c_c is a constant and n_c' is the number of nodes that are present in the probing paths. Furthermore, we show that if a small fraction of error, $\delta_c > 0$ can be tolerated, then $m_{c,\delta_c} \leq m_c$.

(e) Since congestion can occur much more frequently than failures, we devise a methodology to accurately localize congested nodes in real-time using active probing. Simulations are performed on networks of various sizes and under various probing scenarios and are compared against the fundamental limits. The monitoring approach requires just one bit of payload in the probe packet for congestion localization; hence, it does not overburden the network. The inference of the nodal states is performed using a message-passing algorithm that is known to have a computational complexity proportional to $O(d \log d)$, where d is the block length of the low density parity-check (LDPC) code. The computational complexity of congestion localization is proportional to $O(n_c' \log n_c')$.

The next part of the thesis investigates automatic techniques for quantitative analysis of biological specimens imaged using electron tomography (ET). ET is a general method for three-dimensional reconstruction of electron transparent objects from a series of projection images recorded with a transmission electron microscope. It is an emerging tool to describe the three-dimensional (3D) architectures of large molecular complexes, viruses, and cells. In biological applications, electron tomograms typically contain rather signal-to-noise ratios (SNR). In electron tomograms imaged at cryogenic temperatures, this stems in part from the use of low electron doses, which help minimize radiation damage to the specimen. Tomograms collected from sections of plastic-embedded specimens can be obtained at higher electron doses to improve the signal, but they contain electron dense stains that can distribute unevenly and contribute to the background, adding considerable noise to the tomogram. Image interpretation of both kinds of tomograms is complicated further since most data collection schemes result in a “missing wedge” of information that results in anisotropic resolution and degradation of image quality in the reconstructed tomogram in the direction of the electron beam.

The challenge of achieving high resolution with electron imaging of unstained specimens is that doses that are high enough to get a good SNR lead to unacceptable specimen damage, while doses that are low enough to preserve the specimen lead to poor image quality. Hence, there is a trade-off between the dosage level and the damage caused to the specimen. As techniques for data collection and tomogram reconstruction are becoming more streamlined, the development of methods to analyze the enormous amounts of information in these tomograms remains a major challenge [23]. A fundamental step in analyzing such datasets is to employ robust methods for 3D segmentation that can work well at low SNRs in an automated fashion. The target of these strategies is to improve the signal as much as possible relative to the noise level to mine 3D information from complex tomograms. A long-term goal of this type of electron tomography is to interpret the spatial arrangement of the constituent molecular and cellular components at high resolution and facilitate the overall goal of speeding up the process of reducing gigabyte-sized tomograms to relevant byte-sized data.

A number tools for quantitative interpretation of tomograms have been developed and applied to biological specimens [24-26]. The data quality of such tomograms can vary widely because of differences in specimen preparation, data collection schemes, or origins of contrast as well as from using different algorithms to align and reconstruct the data into a final 3D volume. Given that any particular denoising algorithm may not perform uniformly well on such diverse datasets, it is essential to filter tomographic data based on analysis of the noise in a given dataset rather than an *a priori* assumption of a particular noise model. A fundamental step towards automated analysis of large amounts of data for statistical inference is to identify robust methods for 3D segmentation that function effectively at low signal-to-noise ratios. Our main focus is to apply information theoretic and machine learning approaches in mining relevant byte-sized information from gigabyte-sized tomograms.

Initially, we present a quantitative evaluation of denoising algorithms for biological tomograms. In particular, we provide a comparative investigation of several image and transform-domain denoising techniques combined with tools for feature extraction from electron tomographic data. Using tomograms collected from stained specimens at room temperature and from frozen hydrated specimens at cryogenic temperatures, we perform a spectrum of denoising

algorithms based on nonlinear anisotropic methods, wavelet-based methods, and filtering-based techniques, without making *a priori* assumptions about the statistics or the type of noise present. We then identify the optimal denoising strategy for each tomogram using quantitative measures such as the goodness-of-fit (GOF) test based on the Kullback-Leibler (KL) distance [27], Fourier ring correlation [28], and single-image SNR [29]. Information is then extracted from denoised test tomograms either by automated "one-click" segmentation in the case of tomograms obtained from stained-plastic embedded specimens of HIV-1 infected macrophages, or by automated feature extraction based on template matching for tomograms obtained from plunge-frozen *Bdellovibrio bacteriovorus* cells. The value of denoising for tomogram interpretation was assessed by qualitative and quantitative comparison of automated feature extraction versus the information retrieved manually by an experienced user from the same tomograms. We demonstrate that this strategy can be used as a high-throughput approach to identify an "optimal" denoising method for a given type of tomogram, and that denoising enables rapid, automated, and accurate segmentation of biologically relevant features of interest.

Next, we present an approach for automatic texture-based segmentation of mitochondria in MNT-1 cells imaged using an ion-abrasion scanning electron microscope. Our method is based on block-wise classification of images into a trained list of regions. This segmentation problem is challenging because of low contrast and signal-to-noise ratio (SNR), appearance, geometry, and viewpoint variation. Given manually labeled images that contain both segmentation maps and rough annotations of regions, our goal is to learn models that can localize novel instances of the regions on test datasets. To facilitate automatic segmentation, we improve the SNR of the tomogram by implementing a two-dimensional texture-preserving filter that incorporates a spatially varying fidelity term, and thereby locally controls the denoising of image regions proportional to their content. Binary classification of mitochondria in the cytoplasm is executed by exploiting the texture features. We investigate region features based on a histogram of pixel textons. Block-wise classification is performed by histogram matching using a *nearest neighbor* classifier and the chi-squared statistic as a distance measure. Segmentation results demonstrate that the proposed approach, using minimal training data, performs close to a semi-automatic approach carried out using a variational level-set method and a manual segmentation by an

experienced user. We then analyze quantitative measures, such as volume of cytoplasm occupied by mitochondria, difference between surface area of inner and outer membranes and mean mitochondrial width, which differentiate a cancerous cell from a normal one. To test the accuracy of our approach, the quantities are compared against manually computed counterparts. Next, we test the robustness of texture-based segmentation on tomograms of *Liposomal Doxorubicin* formulations (*Doxil*), an anticancer drug, imaged at cryogenic temperatures and study radii and volume distributions, which are of great interest in caner treatment

The structure of the rest of the dissertation is as follows. Chapter 2, presents a brief overview of various aspects of network and biological tomography. We introduce network tomography and related work in network diagnosis, and discuss probabilistic graphical models and iterative decoding techniques. Next, we summarize electron tomography procedures, and briefly discuss data collection techniques in electron tomography and the biological significance of the data that are analyzed in this work. In addition, we provide a general outline of denoising techniques and segmentation methods. Chapter 3 investigates our scalable active-probing approach for failure and congestion diagnosis using probabilistic graphical models and presents iterative inference techniques. Chapter 4 presents automated denoising and feature extraction for biological tomograms are presented. Chapter 5 investigates automatic techniques for segmentation and quantitative analysis of mitochondria in MNT-1 cells imaged using an ion-abrasion scanning electron microscope, and tomograms of Liposomal Doxorubicin formulations (Doxil), an anticancer nano-drug, imaged at cryogenic temperatures. Finally, we present conclusions and future work in Chapter 6.

CHAPTER 2

BACKGROUND INFORMATION

This chapter presents background information on network diagnosis, probabilistic graphical models, iterative decoding techniques, electron tomography, data collection procedures for electron tomography, denoising techniques, and segmentation methods. It also provides details about how machine learning and information theoretic methods have been applied to solve problems in network and biological tomography.

This dissertation investigates network diagnosis and automatic techniques for quantitative analysis of biological tomograms using information theoretic and machine learning approaches. The first part of the dissertation studies a general problem of scalability and performance by focusing on two aspects of network monitoring, namely failure and congestion diagnosis. The second part deals with quantitative evaluation of denoising methods and a texture-based automatic segmentation approach for biological tomograms.

Section 2.1 gives a brief overview of the network diagnosis problem and related work in the area of network diagnosis relevant to our work. Section 2.2 provides a concise overview of probabilistic graphical models that is used as a tool to perform inference on large graphs in this work. In Section 2.3, we provide a brief over of iterative decoding techniques, particularly focusing on decoding of Low Density Parity-Check (LDPC) codes, which is the algorithm we apply to perform inference in real-time. In the next three sections, we provide brief introduction to electron tomography, and provide details about the data collection techniques, and the biological significance of the tomographic data that we deal with in this work. In Section 2.7 and Section 2.8, we review some basic denoising and image segmentation techniques. In Section 2.9, we provide details about how information theory and machine learning methods have been applied to network and biological tomography.

2.1 Network Diagnosis

2.1.1 Network Tomography

Studying the internal characteristics of a network using measurements obtained from end-hosts is known as network tomography. The internal characteristics of a network are, for example, connectivity, loss, congestion, and delay at either a node or link level. Researchers have investigated both passive and active probing methods to detect and mitigate network problems. Simple examples of probes to analyze the status of a path include traceroute/pings and bandwidth monitoring tools such as the pipechar [30].

2.1.2 Active vs. Passive Probing

Passive measurements can be obtained using network tools that are built into the routers, such as simple network management protocol (SNMP) [10], Netflow [11], and RMON [12], to monitor the internal status of the network. Typically, routers have to be polled to collect network statistics such as delay, loss, and available bandwidth. To obtain useful monitoring information, a large amount of data has to be collected, which can be prohibitive if the network size is large. For instance, studies have shown that periodic polling of Cisco-4000 series Netflow-enabled routers on a local network decreased the throughput by as much as 15-20% [1]. In an inter-domain setting, the domains are managed by multiple ISPs, and the service providers do not disclose confidential information about their domains. Hence, active probing may be the only convenient choice when the interior of a network is not assessable directly [13], thus avoiding polling from internal nodes. The detection and localization may also be made easier using end-end measurements, since they contain global information about a network. Passive approaches have been investigated in [2], where the authors argue that the passive approach does not generate probe traffic.

2.1.3 Link Failures

Link failures can be caused by problems related to any network component at or below the IP layer, such as fiber cuts, optical equipment malfunctions, protocol misconfigurations, and router hardware/software failures. Router connectivity may be lost as a result of hardware

failures, router processor overloads, software errors, protocol implementation, and misconfiguration errors [3]. Network component failures affect multiple IP links since they share these components. Unplanned outages or scheduled maintenance can also be a contributor to network failures [3]. Routing instability caused by router configuration errors, transient physical and data link problems, and software issues contributes to poor end-to-end performance [31]. Failures may also occur as a result of malicious attacks. Persistent congestion, on the other hand, is a symptom that occurs because of overload [32], traffic clustering, and malicious attacks such as distributed denial of service (DDos) and worm propagation [13].

2.1.4 Nodal Congestion

Congestion occurs much more frequently than failures in a network [33]; it can be considered as a minor form of failure or as a soft failure. Long-term or persistent congestion arising from bandwidth stealing or flooding of malicious traffic causes the legitimate user to experience longer delays, higher loss rates, and lower throughput, and such effects last for a longer period of time. Human intervention is necessary to overcome such types of congestion. The authors in [34] argue that 80% of the congestion occurs within the access ISP network, which is operated at a higher utilization levels than other networks. Furthermore, in the next few years, voice-over-IP (VoIP) will become a cost effective alternative to traditional telephone services. The factors that affect VoIP services are congestion, link failures, and routing instabilities resulting from router/link failures [33]. Failure and congestion diagnosis can help in many traffic engineering problems such as routing protocols, network design, and management.

2.1.5 Related Work in Network Diagnosis

Existing tools identify congested/faulty segments, but are not effective in localizing the actual congested nodes or faulty links [4, 35-37]. Markopoulou *et al.* [3] analyze IS-IS routing updates from Sprint's IP network to differentiate failures in an IP backbone. Another study on Sprint's backbone network [32] has shown that link overload occurs 80% of the time because of link failures; this can be alleviated by deflecting the packets through less utilized alternate paths. Multiple links rarely fail simultaneously on the same path. Hence, locating failures and congestion on different routes would assist in path selection, and improve network path diversity.

Chen *et al* [20] present an algebraic approach to monitor k linearly independent paths that can describe all the $O(n^2)$ paths, where n is the number of end hosts on a overlay network. They show that k grows as O(nlogn) through linear regression tests on synthetic and real topologies. On the other hand, conventional methods of detecting congestion use packet pair correlation [4], where the end-to-end delays or losses are observed at the end hosts. The work in [35] proposes a practical method to infer congested segments in real-time based on an indirect inference methodology using multiple end-to-end measurements. In [36], a tool called *Pathneck* is presented that allows end users to efficiently and accurately locate the bottleneck links on an internet path using a distributed framework. A methodology for the measurement and classification of bottleneck links based on the investigation of links within a managed domain of a carrier ISP or between neighboring carriers is addressed in [37]. An algorithm for identifying component failure in Shared Risk Link Groups (SRLG) based on Bayesian network is studied in [38]. A probing scheme that uses O(nlogn) measurements based on weighted-set cover algorithms is presented in [39], where n is the number of edge routers on an overlay network. Habib *et al*. [13] investigate a monitoring scheme that requires $O(n)$ probes, if the congestion is less than 20% and $O(n^2)$ probes when the congestion is more than 20%. The work in [40] develops a statistical framework based on linear model theory that exploits the fact that network paths share common links.

Machine learning approaches using probabilistic graphical models have been widely studied in image restoration and coding [41, 42]. In networking, failure diagnosis based on event correlation [43], Bayesian belief networks [15] and heuristic algorithms [44] have been explored, but so far the scalability issue has been less studied. The authors in [5] consider scalability of measurement-based network monitoring in a limited setting of multicast probing. In recent work by Wen *et al.* [45], an efficient failure-diagnosis algorithm is studied for all-optical networks based on sequential probing for both single and multiple failures. The focus of the work in [45] is on linear optical networks under noiseless measurement conditions. A related work on failure diagnosis [16] develops an algorithm for active probing and discusses mini-bucket approximations and their diagnosis complexity. A closely related approach is the application of factor graphs to estimate the link loss exploiting reverse multicast trees for data aggregation in

sensor networks where the construction of the graph is based on link costs and path flows, which are assumed to be independent and identically distributed [17].

2.2 Probabilistic Graphical Models

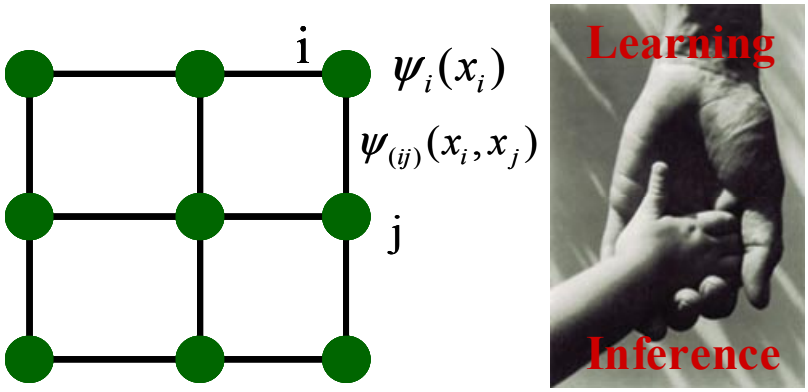


Figure 1. (Left) Example of Markov network and (Right) pictorial representation of learning and inference.

Probabilistic graphical models (PGM) are a marriage between probability theory and graph theory. They have recently gained importance, and hence form one of the backbones of machine learning algorithms. The basic idea of probabilistic graphical models is modularity, which essentially means that complex systems are built by combining simple parts, and probability theory acts as an interface between models and data. PGMs represent the probabilistic relationship between random variables [21], for example, in QMR-DT networks [22] to model symptoms and diseases. They are the main tool for performing inference on large graphs and learning model given a set of examples (data). A pictorial representation of inference and learning is depicted in Figure 1. The two main classes of PGMs are undirected models composed of Markov random fields and Markov networks, as shown in Figure 1, and directed models that include Bayesian/belief networks, as shown in Figure 2 . The former finds application in areas such as image analysis [41, 42], and the latter in causal analysis [22]. The joint probability distribution for Markov network is represented as

$$P(x) = \frac{1}{Z}\prod_i \psi_i(x_i)\prod_{(ij)} \psi_{(ij)}(x_i, x_j), \quad (1)$$

where ψ and Z represent the potential and partition functions, respectively.

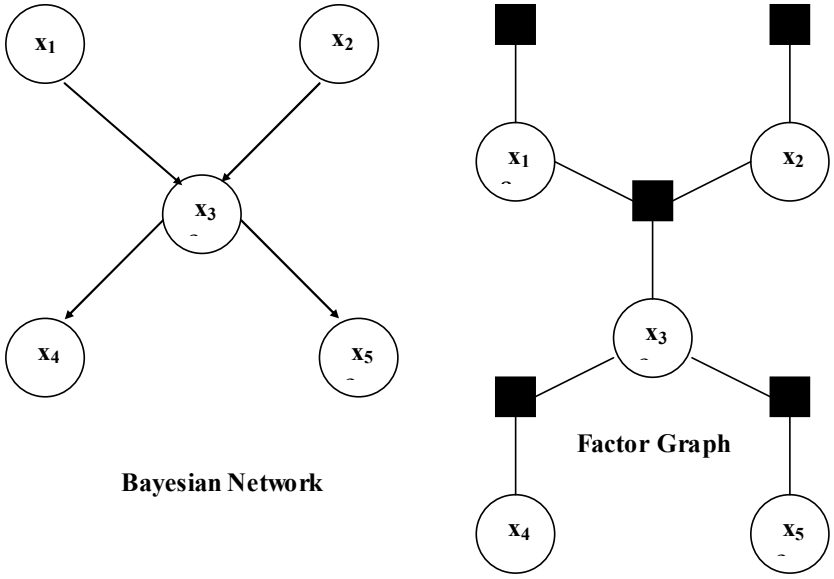


Figure 2. Conversion of Bayesian network to a factor graph.

Next, we show an example as to how a Bayesian network can be converted to a factor graph as in Figure 2. Denoting the set of parents as $pa(x_i)$, we have

$$P(x_1, x_2, ...x_5) = \prod_{i=1}^{5} P(x_i \mid pa(x_i)). \tag{2}$$

For conversion, we simply introduce a function node for each of the conditional probability density functions, $P(x_i \mid pa(x_i))$, as in (2) (shown in Figure 2) and draw edges from this node to x_i and its parents $pa(x_i)$.

2.3 Iterative Decoding Techniques: Low-Density Parity Check (LDPC) Codes

Low-density parity-check codes (LDPC) [46] are block codes characterized by a generator matrix $\boldsymbol{G}$ of dimension $k \times n$, where k is the message length and n is the codeword length [47]. The codeword c is of dimension $1 \times n$, corresponding to a message m of dimension $1 \times k$, together which satisfy the equation

$$c = m\boldsymbol{G}. \tag{3}$$

The code is called a systematic code if the message m appears intact in the codeword c. Hence, after column permutations, the generator matrix $\boldsymbol{G}$ can be written as

$$\boldsymbol{G} = \left[\boldsymbol{I}_{k\times k} \mid \boldsymbol{P}_{n-k\times k}\right], \tag{4}$$

where $\boldsymbol{I}_{k\times k}$ is the identity matrix, and the matrix $\boldsymbol{P}_{n-k\times k}$ corresponds to the systematic portion and the parity bits in the codeword respectively. The block code can also be represented by a parity-check matrix, $\boldsymbol{H}_{n\text{-}k\times k}$, which satisfies $\boldsymbol{HG}^{\boldsymbol{T}} = \boldsymbol{0}$. Therefore, for any codeword,

$$\boldsymbol{Hc}^{\boldsymbol{T}} = \boldsymbol{HG}^{\boldsymbol{T}} m^{T} = 0. \tag{5}$$

The bits of codeword c satisfy the parity-check equations represented by the rows of $\boldsymbol{H}$. The parity-check matrix is given by

$$\boldsymbol{H} = \left[\text{-}\boldsymbol{P}_{k\times n-k} \mid \boldsymbol{I}_{n\text{-}k\times n-k}\right]. \tag{6}$$

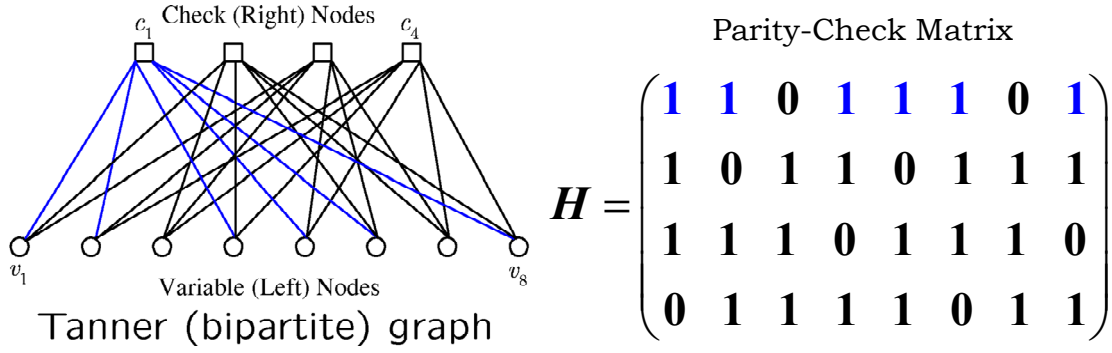


Figure 3. Example of (3, 6) code represented using a Tanner graph and parity-check matrix.

Here, we assume that we are dealing with binary codes and the addition operation is modulo 2. The density of the matrix is given by the ratio of the number of 1's in the matrix to the total number of entries in the matrix; hence, an LDPC code is a code whose parity-check matrix sparse. A $(j,k)-$regular LDPC matrix is shown in Figure 3, which has exactly $k=6$ ones is each row and exactly $j < k$;$j=3$ ones in each column.

The parity-check matrix can also be represented by a bipartite or a Tanner graph as in Figure 3. A bipartite graph has two types of nodes: bit nodes, and check nodes, which correspond to the code bits and the parity check equations (i.e., the rows of $\boldsymbol{H}$), respectively. There exists an edge between the bit and the check node if the code bit corresponding to the bit node participates in the parity check equations corresponding to that check node. Syndrome decoding is one of the

standard methods for decoding block codes, where the syndrome is chosen as a look up index in a table of possible error vectors. This is impractical for practical code lengths because of the large number of possible error vectors. An alternative method of decoding, using a much simpler algorithm based on message passing, assumes that the bi-partite graph corresponding to the LDPC code is cycle-free, and hence optimal. Even in the cases where the code is irregular and the graph is not cycle-free, the optimality is not proved, but the message passing algorithm works significantly well. The message passing decoder works by passing messages between the check and the bit nodes, thus improving the accuracy of its decisions at each iteration.

Next, we present a brief overview of the message passing algorithm assuming a $(j,k)-$ regular LDPC code. Let $\boldsymbol{b}=[b_0,b_1,...b_{K-1}]$ be the transmitted message, where $b_i \in \mathrm{B}$ and B is the transmitted alphabet. The vector $\boldsymbol{b}$ is encoded to codeword, $\boldsymbol{c}=[c_0,c_1,...c_{N-1}]$. Let the received codeword be $\boldsymbol{r}=[r_0,r_1,...r_{N-1}]$. The *maximum a posteriori* (MAP) receiver picks *b* to maximize the *a posteriori* probability $Pr[b_k=b\,|\,\boldsymbol{r}]$ for each *k*, where $b \in \mathrm{B}; \mathrm{B} \in \{0,1\}$. The MAP receiver computes the *a posteriori* log-likelihood ratio (LLR) λ_k given as

$$\lambda_k = \log \frac{Pr[b_k=1\,|\,\boldsymbol{r}]}{Pr[b_k=0\,|\,\boldsymbol{r}]} \tag{7}$$

for eack *k*, and the decision is made as follows,

$$\hat{b}_k = \begin{cases} 0 & \lambda_k \le 0 \\ 1 & \lambda_k > 0 \end{cases}. \tag{8}$$

Using Bayes rule, we can simplify (7) as

$$\lambda_k = \log \frac{f(r_k\,|\,b_k=1)}{f(r_k\,|\,b_k=0)} + \log \frac{\Pr[b_k=1\,|\,\boldsymbol{r}_{i\neq k}]}{\Pr[b_k=0\,|\,\boldsymbol{r}_{i\neq k}]}, \tag{9}$$

where $f(r_k\,|\,b_k=1)$ is the conditional probability density function. The first term in (9) represents the intrinsic information, λ_k^{int}, which is output of the k^{th} channel observation and the second term is the extrinsic information, λ_k^{ext}, for the k^{th} bit, resulting from all the other channel observations. The message passing algorithm computes the extrinsic information in a computationally efficient

way. The LLR at the i^{th} iteration converges to the true *a posteriori* LLR shown in (9). The LLR variables are initialized to the intrinsic values:

$$\lambda_k^0 = \lambda_k^{\text{int}} = \log \frac{f\left(r_k \mid b_k = 1\right)}{f\left(r_k \mid b_k = 0\right)}. \tag{10}$$

Let the set of check nodes connected to bit node *i* and the set of bit nodes connected to the check node be represented by the indices *P(i)* and the *Q(i)*, respectively. The cardinality of *P(i)* and *Q(i)* are $j \forall i$ and $k \forall i$. Let the message from the i^{th} bit node to the check node at the n^{th} element in *P(i)* be denoted by $\pi_{in}^{b \to c}$, and let the message from the i^{th} check node to the bit node at the n^{th} element in *Q(i)* be denoted by $\pi_{in}^{c \to b}$. Let the n^{th} element of *P(i)* be *z*, and let $\Pi_{iz}^{c \to b}$ be the set of messages $\forall n \in P(i) \setminus \{z\}$, where $\forall n \in P(i) \setminus \{z\}$ are all the elements of *P(i)* excluding *z*. Therefore, check node *z* is the n^{th} check node connected to the i^{th} bit node. The message from the i^{th} bit node to the z^{th} check node is

$$\pi_{in}^{b \to c} = \log \frac{Pr\left[b_i = 1 \mid \Pi_{iz}^{c \to b}, \lambda_k^{\text{int}}\right]}{Pr\left[b_i = 0 \mid \Pi_{iz}^{c \to b}, \lambda_k^{\text{int}}\right]}, \tag{11}$$

which is the LLR computed at the i^{th} bit node, taking into account input along all the edges, excluding the n^{th} edge. If we assume that the graph is cycle-free, we obtain,

$$\pi_{in}^{b \to c} = \lambda_i^{\text{int}} + \sum_{x \in P(i) \setminus \{z\}} m_{xy}^{c \to b}, \tag{12}$$

where the y^{th} element in *Q(x)* is *i*; hence, the message is sum of the intrinsic information and the summation of all incoming check-to-bit messages, except for that check node for which it is computed. In the first iteration, the check-to-bit messages are initialized to zero, hence

$$\pi_{in}^{b \to c} = \lambda_i^{\text{int}} \forall i \text{ and } n \in P(i). \tag{13}$$

The LLR for the i^{th} bit on the k^{th} iteration is given as,

$$\lambda_i^k = \lambda_i^{\text{int}} + \sum_{x \in P(i)} m_{xy}^{c \to b}. \tag{14}$$

Similarly, let $\Pi_{iz}^{b \to c}$ be the set of messages $\forall n \in P(i) \setminus \{z\}$. The message from the i^{th} check node to the z^{th} bit node is

$$\pi_{in}^{c\to b} = \log \frac{Pr\left[b_z = 1 \mid \Pi_{iz}^{b\to c}\right]}{Pr\left[b_z = 0 \mid \Pi_{iz}^{b\to c}\right]}. \tag{15}$$

Unlike the bit node computation, for the check node,

$$\pi_{in}^{c\to b} = -2\tanh^{-1}\left(\prod_{x\in Q(i)\setminus\{z\}} \tanh\left(\frac{-\pi_{xy}^{b\to c}}{2}\right)\right)\log \frac{Pr\left[b_z = 1 \mid \Pi_{iz}^{b\to c}\right]}{Pr\left[b_z = 0 \mid \Pi_{iz}^{b\to c}\right]}, \tag{16}$$

where the y^{th} element in *P(x)* is *i*. Here again we exclude the edge of interest. The check-to-bit messages contain the extrinsic information, and each incoming messages incident on the i^{th} bit has information about $b_i \in \mathrm{B}$ based on *k-1* observations in the set $\boldsymbol{r}_{z\neq i}$. Therefore, after the first iteration, we have information about $b_i \in \mathrm{B}$ from *j(k-1)* elements out of the possible *N-1*, and at the end of the second from $j(k-1)(k-1)^2$ elements of $\boldsymbol{r}_{z\neq i}$. Eventually, as a function of iteration, the sum of the check-to-bit messages will contain extrinsic information of (9) based on all the elements of $\boldsymbol{r}_{z\neq i}$. But, when the graph contains cycles, the message-passing decoder provides an approximate MAP output, and can be shown to perform significantly well [48].

We now consider an example that demonstrates the operation of the message-passing algorithm picked from [49]. Consider an (8, 4) product code with $d_{\min} = 4$ comprised of a (3, 2) single parity check code with $d_{\min} = 2$ along rows and columns as shown below.

C_0	C_1	C_2
C_3	C_4	C_5
C_6	C_7	

The parity check equations are

$$\begin{aligned} C_2 &= C_0 + C_1 \\ C_5 &= C_3 + C_4 \\ C_6 &= C_0 + C_3 \\ C_7 &= C_1 + C_4 \end{aligned}, \tag{17}$$

and the corresponding parity-check matrix is

$$H = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}. \quad (18)$$

Note that this code is neither low-density nor regular, and hence is a good candidate for demonstrating the message passing algorithm. The Tanner graph for this example is shown in Figure 4.

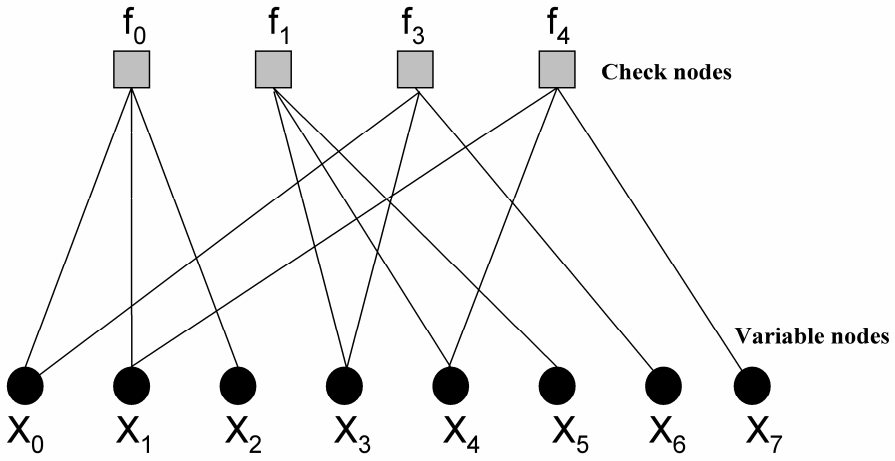


Figure 4. Tanner graph for the example

The code that we choose is:

C_0=1	C_1=0	C_2=1
C_3=0	C_4=1	C_5=1
C_6=1	C_7=1	

and we define $X_i = (-1)^{C_i}$, which gives us $\bar{X}$ as follows:

-1	+1	-1
+1	-1	-1
-1	-1	

The received codeword $\bar{Y}$ is given below. We can observe the sign errors in Y_0 and Y_4.

+0.2	+0.2	-0.9
+0.6	+0.5	-1.1
-0.4	-1.2	

Given $Y_i = X_i + N_i$, where $N_i \sim N(0,\sigma^2)$, and $\Pr(X_i = +1) = \Pr(X_i = -1) = 1/2$, we have

$$\Pr(X_i = x \mid Y) = \frac{1}{1+e^{-2yx/\sigma^2}}. \tag{19}$$

Step 1: Initialize

$$\begin{aligned} q_{ij}(1) = P_i &= \Pr(X_i = -1 (C_i = 1) \mid Y_i) = \frac{1}{1+e^{-2y_i/\sigma^2}} \\ &= (0.31,\ 0.31,\ 0.97,\ 0.083,\ 0.12,\ 0.99,\ 0.83,\ 0.99) \end{aligned}, \tag{20}$$

and $q_{ij}(0) = 1 - P_i \ \forall i, j$.

Step 2: $\pi_{ji}(0) = 0.5 + 0.5 \prod_{i' \in R_j \setminus \{i\}} (1 - 2q_{i'j}(1))$ and $\pi_{ji}(1) = 1 - (0.5 + 0.5 \prod_{i' \in R_j \setminus \{i\}} (1 - 2q_{i'j}(1)))$.

Hence, in this case,

$$\begin{aligned} \pi_{00}(0) &= 0.5 + 0.5 \prod_{i' \in \{1,2\}} (1 - 2q_{i'0}(1)) \\ &= 0.5 + 0.5(1 - 2(0.31))(1 - 2(0.97)) = 0.32 \end{aligned}, \tag{21}$$

$\pi_{01}(0) = \pi_{00}(0)$, and $\pi_{02}(0) = 0.5 + 0.5(1 - 2(0.31))(1 - 2(0.31)) = 0.57$. Thus, all $\pi_{ji}(0)$ can be obtained.

Step 3: Computation of q_{ij}. $\tilde{q}_{00}(0) = (1 - P_0) \prod_{j' \in \{2\}} \pi_{i'j}(0) = (1 - 0.31)(0.22) = 0.15$ and

$\tilde{q}_{00}(1) = 0.24$. Hence, $q_{00}(0) = \frac{\tilde{q}_{00}(0)}{\tilde{q}_{00}(0) + \tilde{q}_{00}(1)} = 0.38$ and $q_{00}(1) = \frac{\tilde{q}_{00}(1)}{\tilde{q}_{00}(0) + \tilde{q}_{00}(1)} = 0.62$,

respectively.

Step 4: $Q_i(0) = K_i(1 - P_i) \prod_{j' \in C_i} \pi_{ji}(0)$ and $Q_i(1) = K_i P_i \prod_{j' \in C_i} \pi_{ji}(1)$, where the constants K_i are chosen such that $Q_i(0) + Q_i(1) = 1$.

Step 5: Termination and decoding:

$$\hat{c}_i = \begin{cases} 1 & if \ \ Q_i(1) > 0.5 \\ 0 & else \end{cases} \forall i, \tag{22}$$

The termination is reached if $\hat{c}\boldsymbol{H}^{\boldsymbol{T}} = 0$ or the number of iterations has reached the decided maximum. In the above example, the decoder converges to the correct codeword in seven iterations. At the 7^{th} iteration the values of $Q(1) = (.7399, .3381, .9692, .4086, .7869, .9567,$

.7754, .9923) , which is decoded as $c = [1\ 0\ 1\ 0\ 1\ 1\ 1\ 1]$. This is the original transmitted codeword.

2.4 Electron Tomography

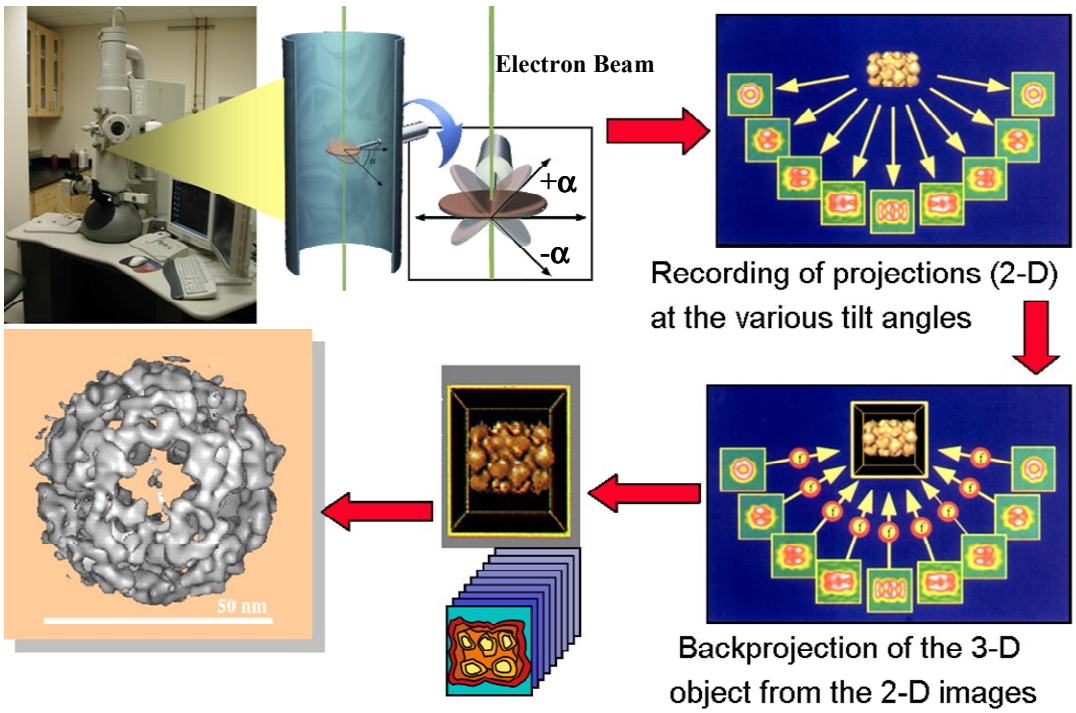


Figure 5. Data collection and reconstruction procedures ([50])

In electron tomography, efforts to establish tools for quantitative interpretation of tomograms are beginning to be applied to a range of biological problems, as reviewed recently in [51]. At high doses, electrons interact with unstained organic matter, which leads to the breaking of chemical bonds and the creation of free radicals, which in turn cause further secondary damage. At low doses it is possible to scatter electrons from biological matter such as proteins and record molecular images, but at the expense of poor signal-to-noise ratio (SNR) [23]. The data collection includes taking a series of images at different relative tilts of the beam and specimen, as shown in Figure 5. The electron microscope provides 2D projection images. Tomograms are generated by the computational analysis of a series of electron micrographs taken at varying tilt angles, usually from -70° to +70°. Using algorithms that implement weighted back

projection or the simultaneous iterative reconstruction technique (SIRT), one can obtain 3D volumes of the specimen from projection images, as shown in Figure 5.

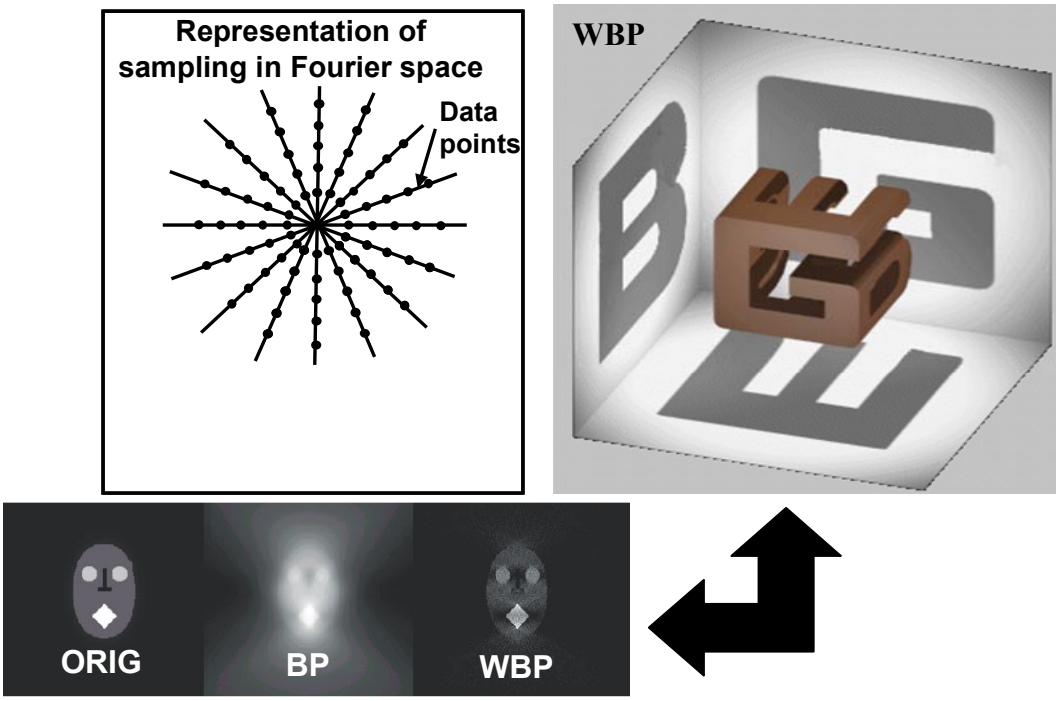


Figure 6. Sampling in Fourier space and an example of reconstruction using back projection and weighted back projection.

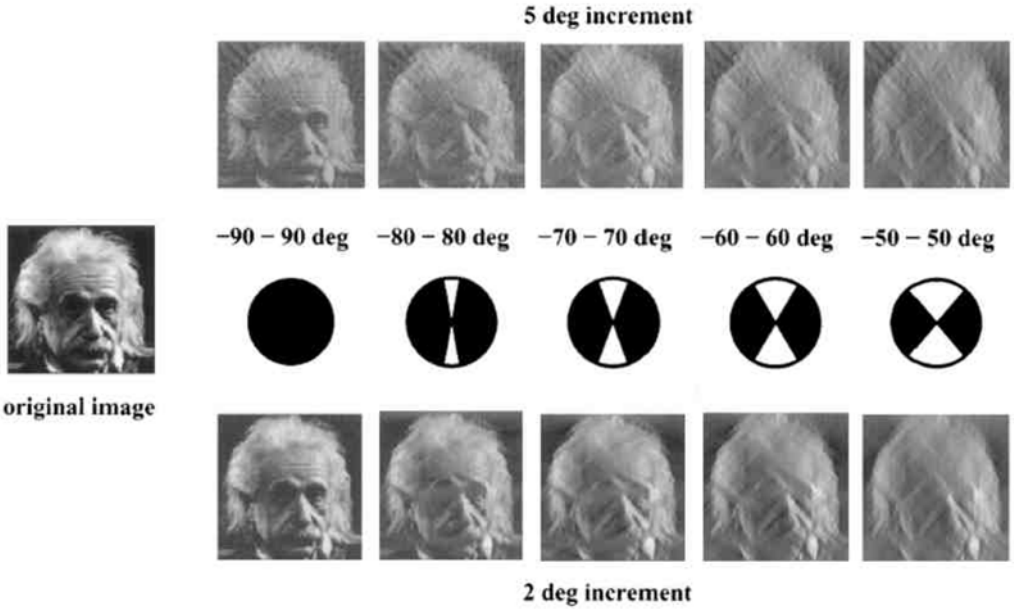


Figure 7. Pictorial representation of a "missing wedge" of information in reconstruction ([50])

Figure 6 shows a pictorial representation of sampling in Fourier space during reconstruction using WBP, where high frequencies are under sampled. The figure also shows an example of reconstruction using back projection (BP) and WBP; the WBP reconstruction is better than the BP reconstruction. In ET, the range of tilt angles is often limited to ±70° because of the physical constraints imposed by the microscope, sample, and holder, which leads to a 'missing wedge' of information, as represented in Figure 7. This loss of information results in artifacts in the reconstruction, including an elongation of the reconstruction in the direction of the optic axis, perpendicular to the tilt axis (anisotropic resolution), and poor image quality in the reconstructed tomogram in the direction of the electron beam. All reconstructed volumes contain a significant level of noise that arises in part from the inherent limitations of current electron optics, such as chromatic and spherical aberration, beam incoherence, point spread, and certain terms of the lens transfer function, such as contrast transfer and inelastic scattering (although the latter can be reduced by an energy filter). Other sources of noise arise from the CCD detector and in part from computational manipulation of the data, including inaccuracies in alignment of the tilt series and reconstruction of objects that have missing information in Fourier space. In the special case of cryo electron tomography, the necessity of using low electron doses is an additional factor contributing to noise. The use of dual axis tomography, in which two tilt series are recorded with mutually perpendicular tilt axes, reduces the missing wedge into a 'missing pyramid' [52]. We implement an algorithm based on SIRT that is performed by, at each iteration, comparing re-projections of a reconstruction with each projection (image) obtained in the microscope, using the original projection series as a constraint on the output of the next iteration [53]. This procedure is known to reduce the reconstruction artifacts arizing from noisy projection images. Specific details of data collection procedures for tomograms recorded at both room and cryogenic temperatures are provided in [25, 54, 55].

2.5 Data Collection Procedures for Electron Tomography

This section provides a brief overview of data collection procedures for electron tomography. We particularly focus on data collection at room and cryogenic temperatures, and

present a brief description of ion-abrasion scanning electron microscopy (IA-SEM) used to image large mammalian cells at nanometer resolution.

2.5.1 Data Collection at Room Temperature

Projection images were recorded from 90-nm-thick sections of fixed, stained, plastic-embedded HIV-infected monocyte-derived macrophages through a range of ± 70°, at 2° intervals from 0° to ±50° and 1° intervals from ±50° to ±70°, on a Tecnai 12 Electron Microscope (FEI, Netherlands), at 1μm defocus. The sample was rotated 90° and an orthogonal tilt series of the same area was collected. Prior to acquisition, both surfaces of the sections were coated with 15-nm gold fiducial markers for alignment. Images were recorded on a 4K×4K Gatan CCD camera binned to 2K×2K, at a magnification that produces a pixel size corresponding to 0.44 nm at the specimen level. Acquisition was automated using the Xplore3D software (FEI, Netherlands). Each tilt series was aligned using the IMOD package [56]. The aligned series were then reconstructed in either IMOD, using weighted back projection (WBP), or Inspect3D (FEI, Netherlands), using the Simultaneous Iterative Reconstruction Technique (SIRT) with 17 iterations, to obtain orthogonal tomograms. The orthogonal WBP tomograms were combined in IMOD to obtain the final dual-axis WBP tomogram, and the orthogonal SIRT tomograms were combined in IMOD to obtain the final dual-axis SIRT tomogram. Dual-axis tomography was used because single-axis reconstructed volumes contain a relatively large missing wedge of information in Fourier space, resulting in anisotropic resolution and greater distortion of the 3-D structures of the reconstructed objects.

2.5.2 Data Collection at Cryogenic Temperatures

2.5.2.1 Frozen-hydrated Bdellovibrio Cells

Preparation and analysis of specimens of intact, frozen-hydrated *Bdellovibrio* cells was carried out as described in [57]. Briefly, an aliquot of the cell suspension (3μl) was laid on a thin layer of holey carbon supported on a 3 mm-wide copper grid (Quantifoil MultiA, Micro Tools GmbH, Germany). Excess liquid was blotted off with filter paper after 1 min of incubation, and the grid was plunge frozen in liquid ethane cooled by liquid nitrogen. Specimens were maintained

and imaged at liquid nitrogen temperatures in a Polara microscope (FEI Corp., OR, U.S.A.) equipped with a field emission gun operating at 300 kV. A series of low dose (1-2 e^-/A^2) projection images of the frozen specimen, tilted over an angular range of ±69° at fixed 3° intervals, was recorded in a 2K×2K CCD camera located at the end of a GIF 2000 (Gatan Inc., Pleasanton, CA, U.S.A.) energy filtering system. The effective magnification was 22500×, equivalent to a pixel size of 0.63 nm at the specimen level, and the applied defocus was 15 μm. Full resolution images were aligned with the aid of colloidal gold fiducial markers deposited on the carbon prior to specimen preparation. Three-dimensional reconstructions were carried out by weighted back projection of aligned, 4×4 binned images. The package IMOD [56] was used for alignment and reconstruction.

2.5.2.2 Frozen-hydrated Doxil

Frozen hydrated Doxil (Tibotec Therapeutics, NJ, U.S.A.) specimens were prepared and analyzed as follows. 2 μl's of an undiluted Doxil suspension (2mg/ml) was applied to a thin layer of holey carbon supported on a 3 mm-wide copper grid (Quantifoil MultiA, Micro Tools GmbH, Germany). The sample was blotted for 1.5 sec with a 1 mm blot offset using a vitrobot (FEI Corp., OR, U.S.A.) operating at room temperature and 85% humidity. The grid was then immediately plunged into liquid ethane cooled to liquid nitrogen temperature. Specimens were maintained and imaged at liquid nitrogen temperatures in a Polara microscope (FEI Corp., OR, U.S.A.) equipped with a field emission gun operating at 300 kV. A series of low dose projection images of the frozen specimen, tilted over an angular range of ±70 ° at fixed 3° intervals, was recorded by a 2K×2K CCD camera located at the end of a GIF 2000 (Gatan Inc., Pleasanton, CA, U.S.A.) energy filtering system. The total dose for the tilt seris was 60 e^-/A^2. The effective magnification was 34000×, equivalent to a pixel size of 0.5 nm at the specimen level, and the applied defocus was 5 μm. Full resolution images were aligned with the aid of colloidal 10 nm gold fiducial markers deposited on the carbon prior to specimen preparation. Tilt series were aligned and reconstructed using Inspect3D (FEI, Netherlands). Simultaneous Iterative Reconstruction Technique (SIRT) with 17 iterations was used for reconstruction.

2.5.3 Ion-Abrasion Scanning Electron Microscopy (IA-SEM)

Conventional optical imaging is limited to a resolution of $\sim 0.1\mu m$. Electron microscopy is used for higher resolution, but this needs preparation of specimen by fixation [58] and embedding combined with sectioning using an ultramicrotome. The increased resolution comes at the cost of low speeds in sample preparation and data collection, which is a major concern for high-throughput imaging that involves serial sectioning.

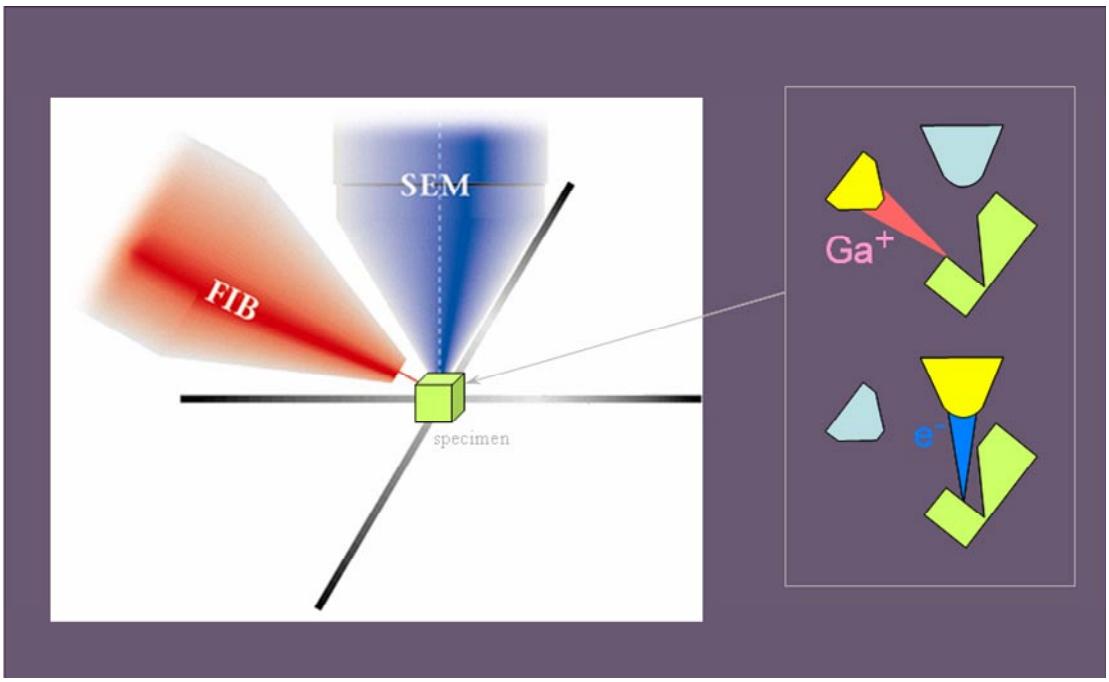


Figure 8. Ion-abrasion scanning electron microscopy (IA-SEM) [58].

To overcome this issue, authors in [58] have implemented a novel dual beam procedure termed ion-abrasion scanning electron microscopy (IA-SEM). This procedure, shown in Figure 8, uses a focused ion beam to create a cut (increments of 20 nm and greater than 100 µm in width) at a designated site in the specimen, after which it is viewed using a scanning electron beam. Iterating these two steps generates surface images of the specimen at regularly spaced intervals which are then put together into a volume representation of the specimen. Recently, to study cell and organelle function at the molecular level, Heymann and coworkers [59] have implemented 3D imaging of large mammalian cells at nanometer resolution. IA-SEM overcomes the problem of the tedious and slow process of 3D volume construction of large eukaryotic cells using serial

sectioning, and has shown that large mammalian cells can be rapidly imaged at resolutions of ~20 nm in the z-direction (direction of section removal) and ~6 nm in the x-y plane (plane of section removal) [59].

The resin embedded melanoma cells were prepared by the procedure described in [59]. The resin blocks were mounted onto standard scanning electron microscope (SEM) stabs and were coated with a few tens of nanometers thick of platinum-palladium layer in the cryo-chamber (Quorum Technologies Ltd) attached to a Nova 200 NanoLab Ion-Abrasion SEM (IA-SEM). Focused ion-beam with current at 7 nA and/or 20 nA was employed to prepare trenches at ~100 μm in width and 40 – 60 μm in length and depth to provide initial cross-sections for searching target cells. Once a target cell was selected, ~1 μm thick of platinum layer was deposited on the top of the interesting area using the Pt gas injector and a focused ion beam, for smooth and conducting surface. The two-dimensional image stacks were collected by alternating electron beam imaging and ion-beam milling, which was controlled by the 'slice-and-view' software (FEI). The scanning electron beam images were collected from the secondary electrons in the immersion lens mode at accelerating voltages of 3 kV and 4 kV, and at beam currents ranging from 68 pA to 340 pA. The cross-sections were milled at inter-image space ranging from 20 nm to 50 nm by a focused gallium ion-beam at an accelerating voltage of 30 kV, and beam currents from 1 nA to 7 nA.

2.6 Biological Significance of Bdellovibrio Bacterium, HIV-1 Virus, Mitochondria and Doxil

Bdellovibrio is small Gram-negative predatory bacteria that invade the periplasm of other Gram-negative bacteria, where they undergo a complex development cycle that terminates in the death of the prey cell [60]. The life cycle of *Bdellovibrio*, in *Escherischia coli*, a typical prey cell, is shown in Figure 9. In their attack-phase, *Bdellovibrios* are propelled by active flagellar motility. With the help of chemotaxis that assists in locating bacteria-rich regions, they randomly collide with prey. In the next stage, they either abort the attachments in seconds or attach permanently within minutes, after which the flagellar motility stops as the bacteria enters the prey

periplasm. In stage 4, *Bdellovibrio* establishes itself in the periplasm and then attaches itself to the cytoplasmic membrane of the prey, and in stage 5, it replicates its DNA, which is called bdelloplast. Then, in stage 6, the *Bdellovibrio* separate into individual cells, and in stage 7, they synthesize flagella. In the final stage, they break the prey membrane and are released to attack other prey Gram-negative bacteria.

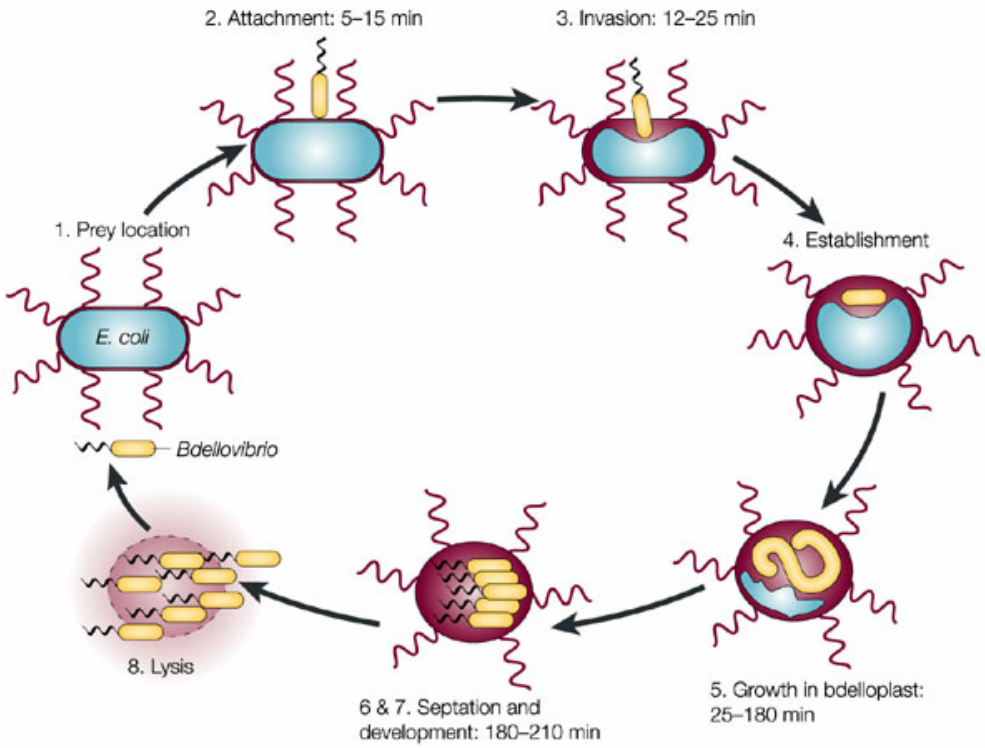


Figure 9. The predatory life cycle of Bdellovibrio (From [60])

We study the spatial distribution of the ribosomal complexes in the *Bdellovibrio* since it provides information on how the protein synthesis takes place and also the transfer of protein to the flagella in stage 7.

The segmentation and study of the volume occupied by the HIV-1 viral spikes are important since they provide us clustering information about the spikes that result in an "entry claw" [61]. The schematic representation of the entry of SIV/HIV-1 virus into the T cells is shown in Figure 10. The "entry claw" is composed of five to seven viral spikes each ~100 angstrom long and spaced laterally ~150 angstrom. The two other modes of viral entry are the

global fusion model (second panel of Figure 10) and the formation of a local pore centered at one of the viral spikes after the entry (third panel of Figure 10). For details, see [61].

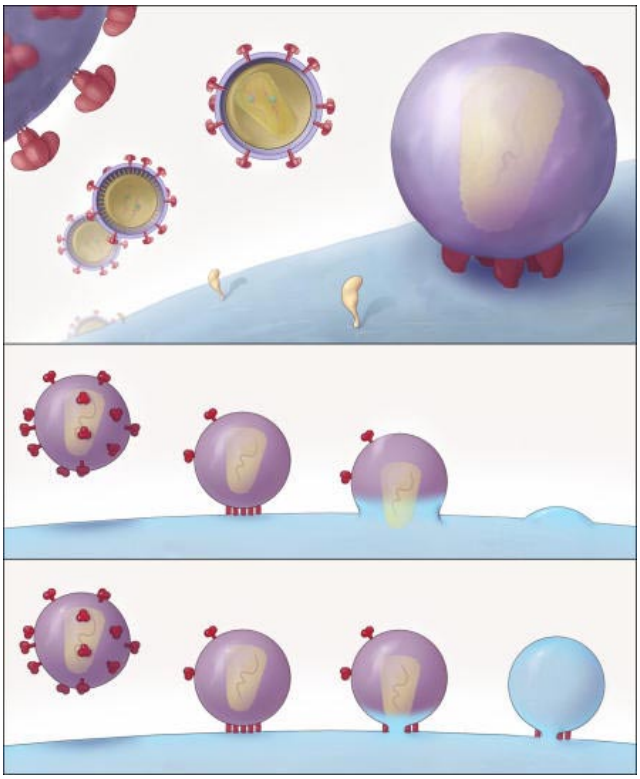


Figure 10. Schematic representation of the entry of SIV and HIV-1 virus into T cells (from [61])

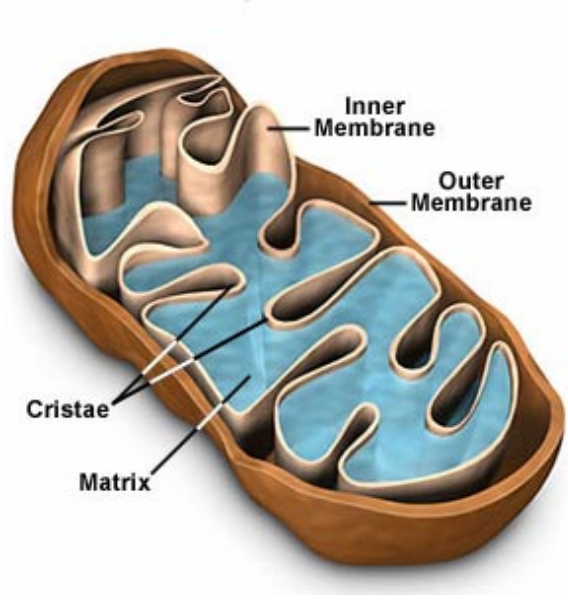


Figure 11. Mitochondria structural features

Mitochondria convert oxygen and nutrients into adenosine triphosphate (ATP) in a complex multi-step process that powers the metabolic activities of a cell through a process known as aerobic respiration. In fact, mitochondria enable cells to produce 15 times more ATP than they could otherwise. The number of mitochondria present in a cell depends upon the metabolic requirements of that cell, and may range from a single large mitochondrion to thousands of the organelles [62]. A mitochondrion is encircled by two specialized membranes as shown in Figure 11 (from [62]), which divides the organelle into a narrow intermembrane space and an internal matrix, each of which contains specialized proteins. The enzymes are able to diffuse through the mitochondrial matrix and act as a catalyst for many of the critical metabolic steps of cellular respiration. The other proteins involved in respiration which, includes the enzymes that generate ATP are embedded within the inner membrane. The protein "porin" forms the outer membrane which, contains many channels and filters out molecules that are too large. The inner membrane is highly convoluted into a large number of infoldings called cristae. It allows only certain molecules to pass through it, and is a bit more selective than the outer membrane. These infoldings increase the surface area for hosting the enzymes for cellular respiration. The inner membrane uses a group of transport proteins that transport only the correct molecules, so only those materials essential to the matrix are allowed in.

Melanin is produced in small organelles known as melanosomes that are transferred, by a currently unknown mechanism, from melanocytes to the surrounding keratinocytes [63]. Melanogenesis is the process through which the pigmented melanin is synthesized in melanocytes [64]. Numerous diseases leading to abnormal pigmentation, in which patients have a compromised immune system, have been documented [63]. The knowledge of intracellular distribution of melanosomes at different stages of biogenesis remains an unanswered question. Structural study, along with the knowledge of the spatial distribution and extent of staining [59], can aid in classifying normal and cancerous cells. Hence, our goal is to automatically segment out mitochondria in MNT-1 cells and extract quantities such as the volume of cytoplasm occupied by mitochondria, the difference between surface area of inner and outer membranes, and mean mitochondrial width, which can segregate a cancerous cell from a normal one.

The anticancer drug doxil (~100 nm) was the first liposomal nanomedicine approved by the Food & Drug Administration (FDA, 1995) [65]. These liposomal formulations of doxorubicin (Doxils) vary in size, structure, and lipid compositions. They have various therapeutic aims [65]. These formulations are expected to accumulate at the disease site and then locally concentrate drug delivery at a controlled rate. These drugs are used to treat both hematological and solid tumors [66]. The microvasculature in tumors is typically continuous, having pore sizes (100-780 nm) large enough for liposomes (100-200 nm) to move from the blood compartment into the extravascular space surrounding the tumor cells [67]. Recent studies on liposomal doxorubicin are focused on issues of acute and chronic toxicity that occur as a result of usage. Furthermore, investigation of the approved formulations of liposomal doxorubicin that are integrated into combination regimes is of great interest [66]. Hence, in this work, we also study the structural properties of doxils such as the radii and volume distribution, which is made possible by texture-based automatic segmentation.

2.7 Denoising Techniques

Denoising algorithms found in image processing literature are based on an assumed noise model or a generic image smoothness model that can be locally or globally applied to an image. All of these algorithms perform remarkably well when the image or the noise model corresponds to the assumptions, but fail in a general setting, resulting in artifacts removing fine structures form the image. Denoising artifacts may include ringing, blur, staircase effects; checkerboard effects; and wavelet outliers. The first approximation most denoising algorithms make is that the noise is additive and can be modeled as

$$y(i,j) = x(i,j) + n(i,j), \tag{23}$$

where $i, j \in I$, $y(i,j)$ is observed noisy pixel value, $x(i,j)$ is true pixel value and $n(i,j)$ is noise perturbation. Two assumptions that are usually true in (23) are that noise is signal independent and noise samples at two distinct pixels are uncorrelated. Thus, the signal-to-noise ratio can be defined as

$$SNR = \frac{\sigma(x)}{\sigma(n)}, \tag{24}$$

where $\sigma(x)$ and $\sigma(n)$ are the standard deviation of the original noise-free image and the noise. The empirical standard deviation of the original noise free image can be obtained as

$$\sigma(x) = \left(\frac{1}{N} \sum_{i,j \in I} \left(u(i,j) - \frac{1}{N} \sum_{i,j \in I} u(i,j) \right)^2 \right)^{1/2}, \tag{25}$$

where N is the size of u. The standard deviation of the noise can be empirically obtained using a known noise model and its parameters. The extent of denoising is dependent on a parameter that is related to the noise variance. All denoising algorithms that are based on (23) assume that the noise is oscillatory and the image is smooth or piecewise smooth, hence, the aim would be to separate the oscillatory part from the smooth parts of the image. The problem with such a model is that many fine structures in an image are oscillatory, and white noise may have smooth components too. Therefore, to overcome such problems, we propose a methodology in Chapter 4 without *a priori* assumption of a particular noise mode. We investigate several standard denoising algorithms based on non-linear diffusion, wavelets, total variation methods, and filtering algorithms and identify an optimal strategy based on quantitative evaluation. In Chapter 5, we apply variational filtering using an adaptive fidelity term that applies different levels of denoising in different regions and preserves textured regions.

2.8 Image Segmentation Methods

Image segmentation using active contours can be broadly classified into two groups namely, parametric [68, 69] and geometric models [70-72]. While parametric models are represented by explicit parameterized curves in a Lagrangian framework, geometric models are characterized by implicit level-sets of two-dimensional functions that evolve over time in an Eulerian framework. The contours of an implicit function defined in a higher dimension are referred to as the level set function (LSF) and are characterized by a zero level set, whose evolution is governed by a partial differential equation (PDE) as a function of time. The advantages of geometric models over parametric models are that topology changes are handled automatically by the inherent breaking and merging process of level set evolution, and since the LSF always remains on a fixed grid, efficient numerical schemes can be devised.

Variational segmentation methods minimize objective functions such as the Mumford-Shah energy functional [73, 74](Chapter 30, [9]):

$$F(I,B)=\int_{\Omega-B}\left[I(x,y)-I_0(x,y)\right]^2 dxdy+\alpha\int_{\Omega-B}\left|\nabla I(x,y)\right|^2 dxdy+\beta\int_B d\sigma, \qquad (26)$$

where I_0 is the original image, Ω is the image domain, I is a simplified image and B is a set of discontinuities (boundaries). The goal is to find the simplified image I and the boundary set B. The term attains a small value when both the original and the simplified images are comparable, and the second term attains a small value when the simplified image is smooth in the areas where there are no boundaries. If the simplified image is piecewise constant away from the boundaries, this term is negligible. The last term measures the length of the boundaries in the image and has a low value if the boundaries are not too "wiggly". It can be supplemented by a regularity term that measures the smoothness (derivatives) of the boundary set. The constants α and β are non-negative values that can be adjusted to favor increasingly smooth functions by increasing α, or by increasing β one can favor small boundary set. Variational level-set formulations are flexible compared to pure PDE driven level-set methods because region based information [75] and shape-prior information [76] can be accommodated into the energy functions, which leads to more robust segmentation results.

2.9 Significance of the Thesis Title

This section provides a brief justification of the thesis title, bridging the gap between the two main topics addressed in this work on network and biological tomography.

The thesis title is "Application of Information Theory and Learning to Network and Biological Tomography" As mentioned in Section 2.1.1, network tomography is a study of the internal characteristics of a network using end-end path-based measurements obtained at end-hosts. The unique role of machine learning is that it provides a framework, based on probabilistic graphical models, to derive theoretical limits and to statistically infer the network states using computationally efficient algorithms. Moreover, variational inference provides a feasible approximation to address the number of spatially dependent measurements needed for diagnosis in large networks, and variational techniques provide a tool to obtain an expression for the scaling

constant that multiplies the growth function. The probes measurements and the links/nodes in the corresponding path form a factor graph.

An example of a sample network topology is shown in Figure 12. The network abstraction is (V,E,W), where V is the set of nodes, E is the set of edges, and W is the set of paths in the logical topology. In the figure, $V=\{1-7\}$, $E=\{a-i\}$, and $W=\{\{a,b,h\},\{e,f,i\},\{a,b,g,i\},\{a,c,d,h\},\{a,c,d,g,i\}\}$.

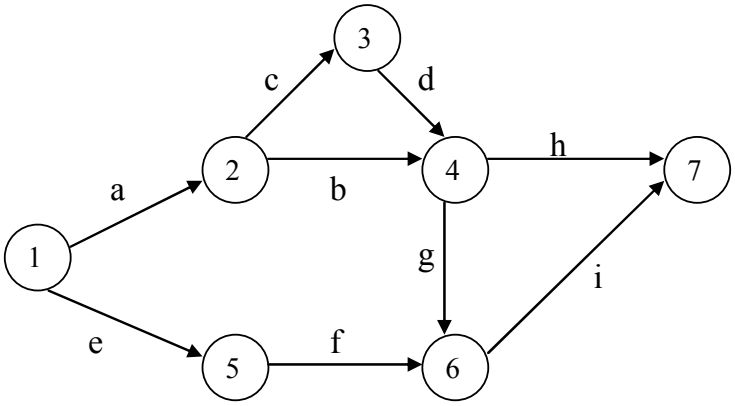


Figure 12. Sample Network Topology

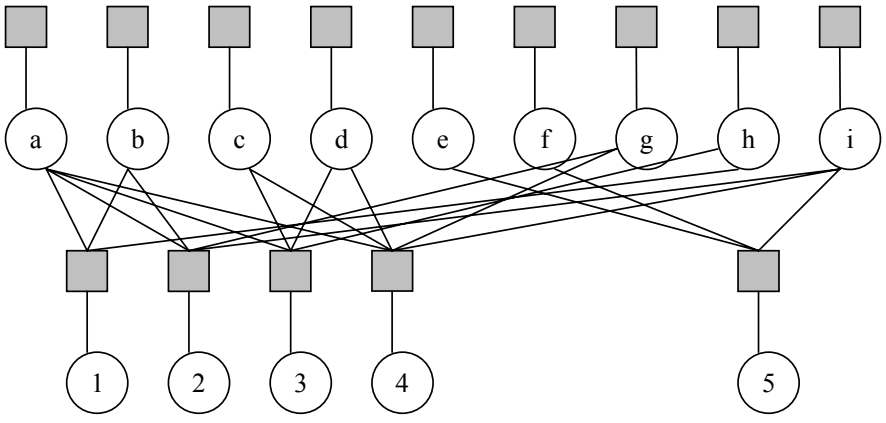


Figure 13. Factor Graph representation of the sample network.

The factor graph representation of the network is shown in Figure 13. It represents a statistical dependence between the link status on one side and the path measurements that includes those links on the other side.

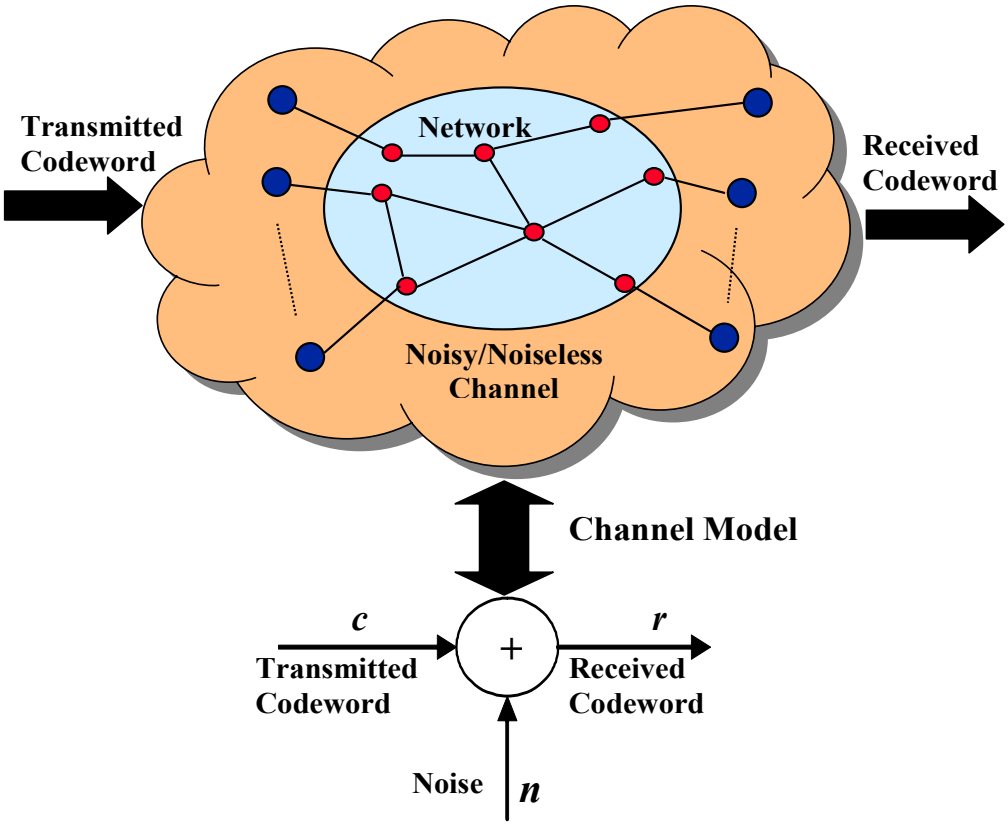


Figure 14. A pictorial representation of the conceptual mapping between network diagnosis and source coding

We employ results from information theory to derive bounds on the average number of measurements needed for accurate diagnosis. From source coding, the minimum number of bits required to encode the message from a discrete alphabet source equals the source entropy. The issue of failure diagnosis or congestion diagnosis is similar to that of the coding problem in the sense that the minimum average number of probes required for network diagnosis is similar to the minimum average number of bits needed to represent the source [45]. The noisy probing scenario is similar to sending bits through a noisy channel, and hence can be mapped to a binary symmetric channel (BSC) with the respective cross-over probabilities. A pictorial representation

of this conceptual mapping is shown in Figure 14. In particular, we can draw the following mapping between network diagnosis and the source-coding problem:

Network states ⇔ Source symbols,

Prior probability of network states ⇔ Prior probability of symbols,

Probe outcomes/path measurements ⇔ Coded symbols through a noisy/noiseless channel,

Average number of probes ⇔ Average code length.

In the case of biological tomography, we apply information-theoretic goodness of fit tests for quantitative evaluation of denoising algorithms and identify the optimal denoising strategy. On the other hand, machine learning provides a plethora of tools and approaches to study automatic classification and segmentation in biological data under low SNR scenarios. Our goal is to learn models that can localize novel instances of the regions on test datasets to facilitate automatic segmentation, given manually labeled images. We investigate region features based on a histogram of pixel textons. Classification of biologically relevant cellular or molecular organelles is executed by exploiting the texture features. Pixel-wise classification is performed by histogram matching using a *nearest neighbor* classifier and a chi-squared statistic as a distance measure.

2.10 Summary

We presented a brief overview of various components of network and biological tomography. Data collection techniques in electron tomography and the biological significance of the data that are analyzed in this work were briefly outlined. The techniques developed in this work can be applied to various applications that involve similar kind of goals. We then previewed shown how information theory and machine learning techniques can be used to solve inference problems in large networks, and generated automated methods for mining relevant byte-sized information from gigabyte-sized tomograms.

The next chapter specifically focuses on the network tomography problem, where we address fault and congestion diagnosis issues.

CHAPTER 3

NETWORK TOMOGRAPHY: FAILURE AND CONGESTION DIAGNOSIS

In this chapter, we study scalability, performance and efficient inference algorithms in the context of unicast path-based measurements. Initially, in Section 3.1, we describe the problem of measurement-based network monitoring using active probing. We provide a general scalability definition, and then study fundamental limits on the relationship between the number of probe packets, size of the network and the ability to diagnose either failed links or congested network elements. Then, we explain the two specific problems that we investigate in this chapter, namely link-failure diagnosis and congestion localization of nodes. In the link diagnosis problem in Section 3.2, variational inference and source-coding techniques are applied to failure diagnosis to derive a lower bound on the average number of measurements with respect to the network size required in the presence of noisy probes outcome. Mapping the scenario of noisy network probing scenario to a binary symmetric channel (BSC) in source coding, allows us to develop a lower bound on average number of measurements. We then validate these bounds using simulation.

Next in Section 3.3, we consider the congestion localization problem as a minor form of failure and derive theoretical bounds on the growth rate of average number of probe measurements by relating the congestion localization in networks using a bit-flipping mechanism to the problem of decoding linear error correcting codes (ECC) over a BSC channel in coding theory. Then, we devise a computationally efficient methodology to accurately localize congested nodes in real-time using belief propagation algorithm.

3.1 Problem Description

Consider a given a network, *G(V,E)*, where *V* and *E* are a set of nodes and links respectively. Let $\boldsymbol{X} = (X_1, ..., X_n)$ be the unobserved status of either nodes or links that obey a certain probability distribution, and $\boldsymbol{Y} = (Y_1, ..., Y_m)$ be a set of *m* observed measurements (i.e.,

samples). Let $P(\boldsymbol{X}=\boldsymbol{x}\,|\,\boldsymbol{Y}=\boldsymbol{y})$ be the conditional probability distribution of $\boldsymbol{X}$ given the measurement set $\boldsymbol{Y}$. Lower-case letters, x_j and y_i denote the realizations, whereas $\boldsymbol{x}=(x_1,....,x_n)$ and $\boldsymbol{y}=(y_1,....,y_m)$ denote a particular assignment of node/link states and m probes outcome respectively. The problem of diagnosis is to obtain the maximum a-priori estimate,

$$\boldsymbol{x}^* = \arg\max_{x} P(\boldsymbol{x}\,|\,\boldsymbol{y}), \tag{27}$$

of the network states given the measurement set $\boldsymbol{Y}$. $\boldsymbol{x}^*$ is the inferred status of the links given observation vector $\boldsymbol{Y}$. Let $P(\boldsymbol{X}\neq\boldsymbol{X}^*(\boldsymbol{Y}))$ be the diagnosis error which measures the difference between the actual and inferred status averaged over a set of measurements. Let δ be a given performance bound. The scalability problem is to obtain the number of measurements needed such that $P(\boldsymbol{X}\neq\boldsymbol{X}^*(\boldsymbol{Y}))\leq\delta$. To be precise, scalability characterizes the growth rate of the minimum set of probes *(m)* as a function of size of the network, *n* (i.e., the number of nodes/links). Link-failure diagnosis and congestion localization of nodes can be considered as two special cases of measurement-based network monitoring are described next.

3.2 Link-Failure Diagnosis

We now consider link failures where the status of link j is denoted as X_j and, $X_j=1$ if the link is working and $X_j=0$ if the link is faulty, for $1\leq j\leq n_f$. n_f is the total number of bi-directional links in the network, i.e., $n_f=|E|$. Thus, the vector $\boldsymbol{X}\in\{0,1\}^{n_f}$ represents the states of all the links in the network. We consider non-ergodic link failures (link-status does not change during the diagnosis process) in this work. Let m_f be the total number of probes that are sent. Let Y_i be the i^{th} probe measurement, sent from a source to a destination, traversing all links on the i^{th} path in the network. $Y_i=1$ if the probe is received, and $Y_i=0$, otherwise, for $1\leq i\leq m_f$. Evidently, a probe is received if and only if all links are working, i.e., Y_i is a logical AND operation, $Y_i=\prod_{j\in\pi_i}X_j$, where π_i represents the parent set of Y_i which consists of all the links on

the path of Y_i for $1 \le i \le m_f$. When measurements are noiseless, $\boldsymbol{x}^*(\boldsymbol{Y})$ obtained from the noiseless probe outcomes, $\boldsymbol{Y}$, represents the most likely estimates.

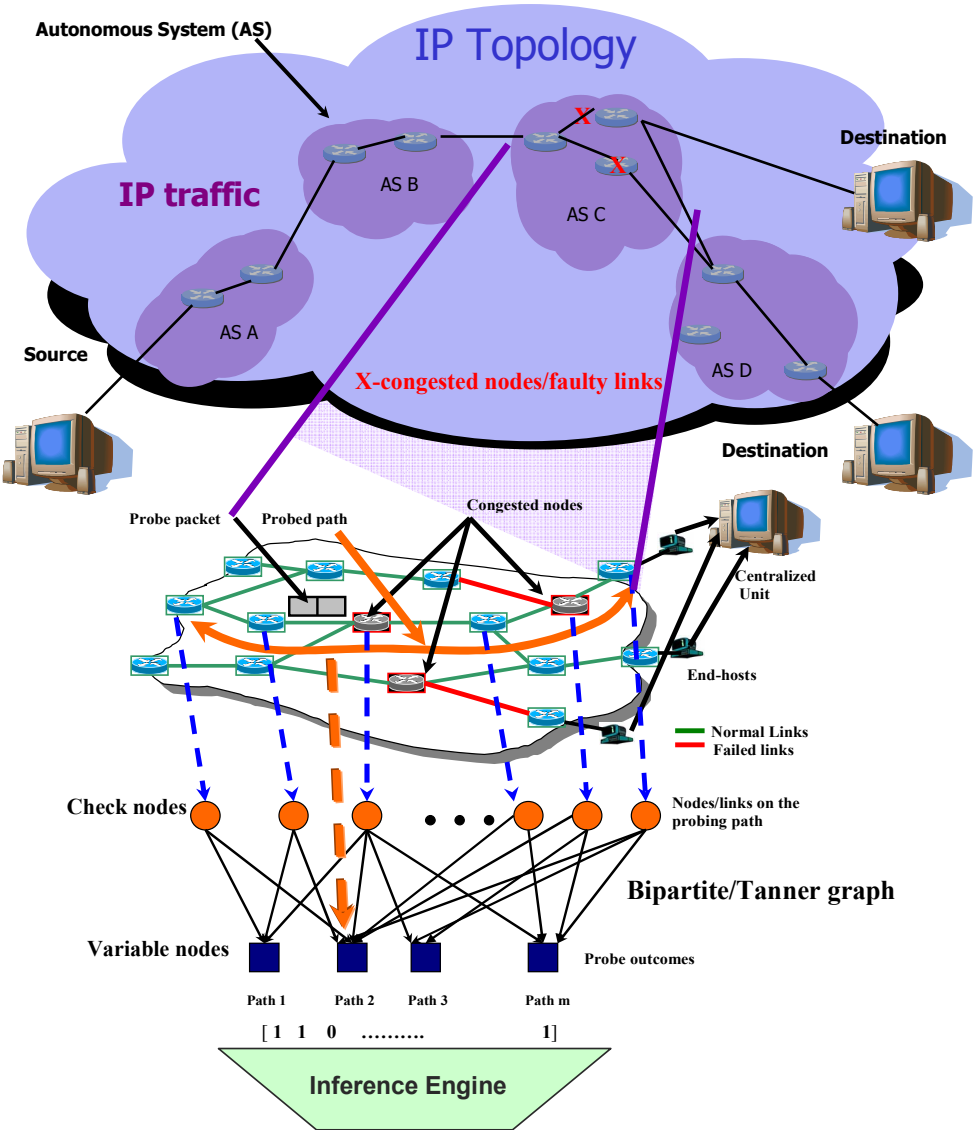


Figure 15. Graphical representation of congestion/faults on the Internet

Measurements are subjected to "noise" and the probe outcomes, $\boldsymbol{Y}$, obtained at the end-hosts do not correspond to the true values of the corresponding probing paths. For example, due to re-routing around failed nodes, a probe may not pass through the links determined by the routing protocol [3]. For instance, the IS-IS routing protocol automatically recomputes alternate

routes in case of IP link failures [3]. In addition, probe outcomes, $(Y_i)_{i=1}^{m_f}$ can experience measurement "noise" (binary inversion) although the links on the routing path are functional. This occurs because of route loops [77] in the routing protocol, which arise from lack of propagation of information about a change in the network connection configuration to every router in the network and ambiguous measurements at the end hosts. In these scenarios, additional probes are needed to infer $\boldsymbol{x}^*(\boldsymbol{Y})$ accurately.

Consider Figure 15, where a percentage of the links are faulty and the status of links $\boldsymbol{X}$ are considered to be random due to uncertain failure and network conditions. $\boldsymbol{Y}$ is also random because of noise in the probe measurements. Then, as shown in the Figure 15, the probing measurements and the links in the corresponding path form a bipartite or a Tanner graph [78]. In the graph, the links in the i^{th} probed path are the parents π_i that influence the end-to-end measurement Y_i. The bipartite graph represents a statistical dependence between the link status and the measurements. The number of probe paths corresponds to the number of measurements m_f. Note that network topology would play an important role in the number of measurements needed. For simplicity, in this work, the probe paths are chosen using greedy heuristics such that they cover a given network topology, and thus results in the structure of the graphical model. The number of measurements can then be obtained from the graph structure. Sampling issues are beyond the current scope of this work and will be dealt as part of future work.

Probe measurements are spatially dependent. For example, two probe measurements are "lost" if they intercept a broken link. Such a spatial dependence can be used to infer states of intermediate nodes/links where measurements are not directly available. Let $g_i = |\pi_i|$, and the computational complexity of exact inference of link status is of $O(2^{\max(g_i)})$, which is inefficient for large networks with a large parent set [22]. A set of most probable link status can be obtained through maximum likelihood *(ML)* estimation from (27). The performance measure of the *ML* estimates in terms of the *most probable explanation (MPE) error* can be evaluated as follows,

$$P(\boldsymbol{X} \neq \boldsymbol{X}^*(\boldsymbol{Y})) = 1 - \sum_{\boldsymbol{y}} P(\boldsymbol{x}^*(\boldsymbol{y}), \boldsymbol{y}). \tag{28}$$

The performance measure can also be represented in terms of bit-error rate *(BER)*, where we

obtain the maximum a-priori estimate $\boldsymbol{x'} = (x_1', ..., x_n')$, where each $x_j' = \arg\max_{x_j} P(x_j \mid \boldsymbol{y})$. In general, it should be noted that that $\boldsymbol{x}^* \neq \boldsymbol{x}'$. $BER = \sum_{j=1}^{n_f} P(X_j \neq X_j'(\boldsymbol{Y}))/n_f$, where $X_j'(\boldsymbol{Y})$ is the maximum a-priori estimate of X_j given the observation vector $\boldsymbol{Y}$. Similarly, as in the case of (28), we can obtain an expression for the *BER* for the j^{th} link as:

$$BER(X_j) = 1 - \sum_{\boldsymbol{y}} P(x_j'(\boldsymbol{y}), \boldsymbol{y}). \tag{29}$$

Given a desired performance bound, the scalability in case of failure diagnosis can be defined as follows. Let $(\delta_{f_1}, \delta_{f_2})$ be the performance bounds, i.e., $P(\boldsymbol{X} \neq \boldsymbol{X}^*(\boldsymbol{Y})) \leq \delta_{f_1}$ and $BER \leq \delta_{f_2}$.

Definition 1: *Scalability is the number of probe measurements m_f that are required to infer the status of the links with the desired performance as the size of the network n_f grows. If the number of measurements needed grows at most linearly with respect to n_f, the diagnostic approach is scalable.*

3.2.1 Lower Bounds on Fault Diagnosis Error

We now assess the number of measurements needed to achieve error such that $P(\boldsymbol{X} \neq \boldsymbol{X}^*(\boldsymbol{Y})) \leq \delta$. A key challenge here is that the measurements are dependent, which invalidates direct applications of the existing results in learning theory. Computing $\boldsymbol{x}^*$ from (27) by brute force, which is needed to calculate the error in terms of the network parameters, is NP-hard [22]. Hence, we resort to bounds on the estimation error. To obtain the bound, we make the following assumptions: (a) the underlying IP topology and the network size are known, (b) the links are assumed to fail independently, and (c) the noise in the probe measurements are *iid*. In most of the earlier work on tomography noise is treated as independent Gaussian random variables [79] or drawn from a known distribution [16, 17]. We do not consider correlated noise in this work. Also, assumption (c) allows us to consider causal independence (different causes

contribute independently to probe failures) in a NOISY-OR Bayesian network. We treat the failure probability as a parameter in the range, $0 \le \rho_f \le 0.5$.

3.2.1.1 Noise Model

The joint distribution of ***X*** and the observation vector ***Y,*** where lower case letters denote the realizations, is

$$P(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}) = \left(\prod_{i=1}^{m_f} P(y_i \mid \boldsymbol{x}) \right) \left(\prod_{j=1}^{n_f} P(x_j) \right) = \left(\prod_{i=1}^{m_f} P(y_i \mid \boldsymbol{x}_{pa(y_i)}) \right) \left(\prod_{j=1}^{n_f} P(x_j) \right) . \quad (30)$$

From Figure 15, the probe outcomes are only dependent on the links in its path, i.e., $P(y_i \mid \boldsymbol{x}) = P(y_i \mid \boldsymbol{x}_{\pi_i})$. Let us denote '+' as the positive evidence, $P(y_i = 1)$ and '-' as the negative evidence, $P(y_i = 0)$ [22]. We consider two types of "noise" in this as scenario described in Section3.2. Noise in the observations are modeled as [22]

$$P(y_i^+ \mid \boldsymbol{x}_{\pi_i}) = (1 - p_{i0}) \prod_{j \in \pi_i} (1 - p_{ij})^{1 - x_j} , \quad (31)$$

where $p_{ij} = P(y_i = 0 \mid x_j = 0); p_{i0} = P(y_i = 0 \mid L)$ and hence represents NOISY-OR Bayesian network. Here, p_{ij} is the probability that the probe *i* fails given that the link *j* in the path fails. p_{i0} is the leak probability, where $L = \{x_j = 1; j \in pa(Y_i)\}$, i.e., the probe fails although all the links in the network are operational. This binary inversion phenomenon occurs at the end-hosts and hence is independent of the route length. The second kind of noise is when the probe arrives at the end-host with a probability $1 - p_{ij}$ even if a single link x_j on its path fails. The re-routing noise, $1 - p_{ij}$ is a function of the route length and hence depends on the physical network topology unlike the measurement noise, p_{i0}. Note that, when route around failures occur most of the links on the native path are retained so that further route loops can be avoided [3, 31, 80]. We let $\theta_{ij} = -\log(1 - p_{ij})$ [22] for mathematical convenience.

3.2.1.2 Most Probable Explanation (MPE) Error Evaluation

Exact MPE: We now derive an expression for the *exact MPE*. From equation (30), $\prod_{j=1}^{n_f} P(x_j) = \prod_{j=1}^{n_f} (1-\rho_f)^{x_j} (\rho_f{}^{(1-x_j)})$ and

$$\prod_{i=1}^{m_f} P(y_i \mid x) = \prod_{i=1}^{m_f} \left[P(y_i^+ \mid \boldsymbol{x}_{\pi_i}) \right]^{y_i} \left[P(y_i^- \mid \boldsymbol{x}_{\pi_i}) \right]^{1-y_i}. \tag{32}$$

Plugging in the conditional densities in (32) and maximizing with respect to $\boldsymbol{x}$ and by using equation (28) we have the expression for *exact MPE*:

$$\max_{x_j} \prod_{i=1}^{m_f} \left[(1-p_{i0}) \prod_{j \in \pi_i} (1-p_{ij})^{1-x_j} \right]^{y_i} \left[1-(1-p_{i0}) \prod_{j \in \pi_i} (1-p_{ij})^{1-x_j} \right]^{1-y_i}. \tag{33}$$

Note that this is computationally intractable for large networks.

Variational MPE: We now derive a lower bound on the *MPE error* using variational inference technique. For any concave function $f(x)$ the tangent plane serves a bound; $f(x) \le \xi^T x - f^*(\xi)$ [22]. The parameter ξ is known as the variational parameter. The bound is better for some values of the variational parameters than for others, and for a particular value of ξ the bound is exact, $f(x) = \min_{\xi} \{ \xi^T x - f^*(\xi) \}$ [22]. A duality exists between $f(x)$ and the concave function $f^*(\xi)$. For simplicity, we assume that the network is homogenous, i.e., $1-p_{ij} = 1-p$ and $p_{i0} = p_0$.

Theorem 1: The lower bound on the MPE error using variational method is given as

$$1 - \sum_{\boldsymbol{y}} \max_{\boldsymbol{x}} P(\boldsymbol{X}=\boldsymbol{x}, \boldsymbol{Y}=\boldsymbol{y}) \ge 1 - \sum_{\boldsymbol{y}} \phi_f{}^{n_f} \left(\prod_{i=1}^{m_f} e^{\beta_i y_i} e^{\alpha_i} \right), \tag{34}$$

where $\phi_f = \max_j(\max(\rho_{f_j}, 1-\rho_{f_j}))$, $\alpha_i = \xi_i(\theta_0 + \theta r_i) - f^*(\xi_i)$, $\theta_0 = -\log(1-p_0)$, $\theta = -\log(1-p)$ are the noise parameters, r_i *is the route length of the* i^{th} *path*, $\xi_i = e^{-(\theta_0+\theta r_i)}/(1-e^{-(\theta_0+\theta r_i)})$ *is the optimal variational parameter,* $f^*(\xi_i) = -\xi_i \log(\xi_i) + (1+\xi_i)\log(1+\xi_i)$, *and* $\beta_i = -(\theta_0 + \alpha_i)$.

The proof is given in *Appendix A*. In (34) $\phi_f{}^{n_f}$ is dependent on the network topology *(n_f)* and failure probability whereas the term inside the parentheses is function of the path length, number

of paths, and the noise parameters. Under noiseless condition, equation (34) is just a function of the network topology *(n_f)* and failure probability since the term $\left(\prod_{i=1}^{m_f} e^{\beta_i y_i} e^{\alpha_i}\right)$ simplifies to a constant as shown in (36). The maximum and minimum values for ξ_i are 0.1179 and 0 respectively for the valid parameter ranges of noise values ($0 \le 1-p, p_0 \le 1$) and route lengths ($3 \le r \le 18$) [81].

Next, we consider a special case by imposing an additional constraint that the path lengths are the same to obtain the closed form expression relating the number of measurements to the number of links. In general, we can design fixed or variable length probing paths using source routing on any given topology. In the case of fixed route lengths, we only select path lengths of a certain fixed length to cover the network once the shortest paths have been obtained. The typical examples of standard topologies where this constraint is valid are star, symmetric tree and dumbbell topology. Under the above constraint, we have the following result.

Corollary 1: *The lower bound on the number of observations for asymptotic error free diagnosis is*

$$m_f / n_f \ge \frac{\log(1-\delta_{f_1})/n_f - \log(\phi_f)}{(e^{-\theta_0} + e^{\alpha})} = \frac{\log(1-\delta_{f_1})/n_f - \log(\phi_f)}{\log\left[(1-p_0) + (1-p_0)^{-\xi}(1-p)^{-r\xi}\xi^{\xi}(1+\xi)^{-(1+\xi)}\right]}. \quad (35)$$

The proof is given in *Appendix B*.

Result 1: Noise-free case- In the absence of noisy observations ($1-p, p_0 = 0$), the lower bound on the number of observations is

$$\frac{m_f}{n_f} \ge \left(\log(1-\delta_{f_1})/n_f - \log(\phi_f)\right)/\log(2). \quad (36)$$

From (36), when the failure probability are equal to 0.5, which is a worst-case scenario, i.e., $\phi_f = 0.5$ then, $m_f \ge n_f$ for $\delta_{f_1} = 0$. Hence, we require the number of observations to be equal to the number of nodes on the bipartite graph, i.e., $m_f = n_f$ for the *MPE error* to be equal to zero asymptotically. In the case when both the noise parameters are large, i.e., $p_0 = 1-\varepsilon_1$,

$1-p=1-\varepsilon_2(r)$, where $\varepsilon_2(r)$ is a function of the route length and $\varepsilon_1, \varepsilon_2(r) \ll 1$ are negligible, we have the following result.

Result 2: Non-scalable case- If $p_0 = 1-\varepsilon_1, 1-p=1-\varepsilon_2(r)$, and $\varepsilon_2(r) \ll \varepsilon_1$, since $\varepsilon_2(r)$ is assumed to be a multiplicative function of route length and hence depends on n_f, then (35) reduces to non-scalable case, i.e.

$$m_f \geq \left(\log(1-\delta_{f_1}) - n_f \log\phi\right) / \log(1+O(\varepsilon_2(r)) = O(n_f{}^{\mu}); \mu > 2 . \quad (37)$$

Note that larger the network size, larger will be the route length which implies that if each link has a noise value ε_2, the denominator of (35) is a function of the longest route, $(\varepsilon_2)^{r_{\max}}$ for variable length probing and $(\varepsilon_2)^{r}$ for fixed length probing scheme. The reason may be due to frequent route loops and re-routing in the routing protocol.

3.2.1.3 Bit-Error Rate (BER) Evaluation

We now evaluate the error using BER as a performance measure by relaxing the fixed probe route length assumption.

Theorem 2: *The lower bound on the bit-error rate using variational method is given as*

$$BER \geq 1-\phi_f(\eta_0+\eta_1)^{\max_j |ch_j|}, \quad (38)$$

where $\phi_f = \max(\rho_f, 1-\rho_f), \eta_z = \max_{i\in\{1,2,..m\}} \max_{\pi_i} P(y_i = z \mid \pi_i); z \in \{0,1\}, \eta_1 = e^{-\theta_0 - r_{\max}\theta}$, $r_{\max}$ is the maximum route length, $\eta_0 = e^{-\theta_0}$, $\theta = -\log(1-p)$, and $|ch_j|$ is the cardinality of the set of children of link *j*.

The proof is given in *Appendix C.* From the above result, we can see that as prior probability of failure increases for fixed noise parameters and varying probe route lengths characterized by $|ch_j|$, the BER increases as expected. This is shown in Figure 16.

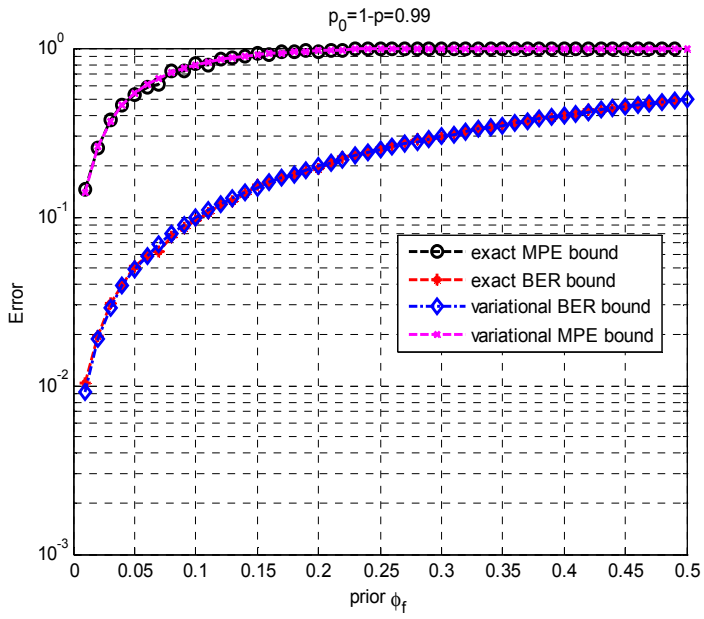


Figure 16. Comparison of the exact *MPE* error/*BER* vs. variational *MPE* error/*BER* lower bounds.

Corollary 2: *If we bound the bit-error rate by the desired performance bound* δ_{f_2}, *the lower bound on the average number of observations for asymptotic error free diagnosis can be given as*

$$m_f / n_f \geq \frac{\log(1-\delta_{f_2}) - \log(\phi_f)}{r_{\max}(e^{-\theta_0} + e^{\alpha})} = \frac{\log(1-\delta_{f_2}) - \log(\phi_f)}{r_{\max} \log\left[(1-p_0) + (1-p_0)^{-\xi^{opt}} (1-p)^{-r\xi^{opt}} \xi^{opt\,\xi^{opt}} (1+\xi^{opt})^{-(1+\xi^{opt})}\right]}, \quad (39)$$

where $\xi^{opt} = \frac{e^{-(\theta_0 + \theta r_{\max})}}{(1 - e^{-(\theta_0 + \theta r_{\max})})}$ and $\alpha = \xi^{opt}(\theta_0 + r_{\max}\theta) - f^*(\xi^{opt})$.

The proof is given in *Appendix D.* Since there are n_f links and m_f observations in the bipartite graph, m_f / n_f denotes the average number of measurements per edge, and hence $\max_j |ch_j| \simeq r_{\max}(m_f / n_f)$. The main difference between equations (35) and (39) is that the assumption on the fixed probe route length is relaxed and (39) is valid for variable probing scenario. Next, if we assume that the route lengths (r) and the number of children of each parent node are equal then we have $\max_j |ch_j| = r(m_f / n_f)$. Hence, the lower bound simplifies in (89) to

$BER \geq 1-\phi(\eta_0+\eta_1)^{r(m_f/n_f)}$. The relation between the two performance measures in terms of the average number of probes per edge required to obtain asymptotic error free diagnosis in the fixed length probing scenario can be given as $\left(m_f/n_f\right)_{BER}=\left(m_f/n_f\right)_{MPE}/r$.

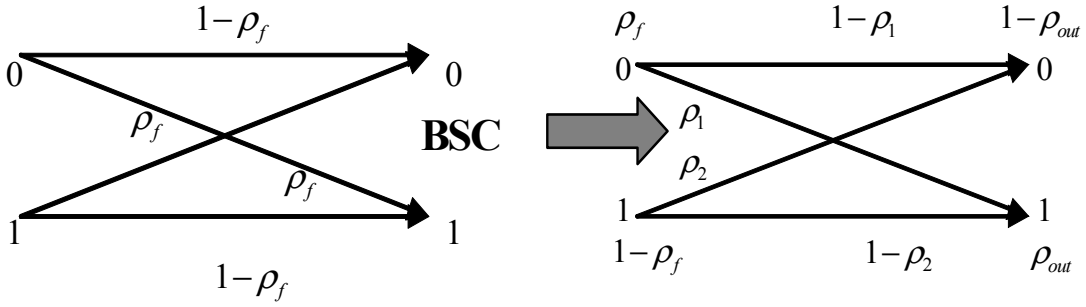


Figure 17. Binary Symmetric Channel (BSC) model for failure diagnosis

3.2.1.4 Entropy Lower Bound on the Fault Diagnosis Error

From source coding, the minimum number of bits required to encode the message from a discrete alphabet source equals to the source entropy. The issue of failure diagnosis is similar to that of coding problem in the sense that the minimum average number of probes required for network diagnosis is similar to the minimum average number of bits needed to represent the source [45]. The noisy probing scenario is similar to sending bits through a noisy channel and hence can be mapped to a binary symmetric channel (BSC) with the respective cross-over probabilities as in Figure 17. Our approach selects the probes that provide the maximum information gain about the network which is essentially reducing the entropy of the network by acquiring information about the network status. Also, the mapping provides a way to analytically obtain the constant that multiplies the growth function. The entropy bound provides a basis for comparing the tightness of the bound in terms of the average number of probes needed for asymptotic error free diagnosis to that obtained by the variational approach in (35). As seen in Figure 18(A-D), the bound on the average number of probes using variational method is not tight for small noise values since the entropy function, *H( )*, provides a much tighter lower bound for this range of noise values.

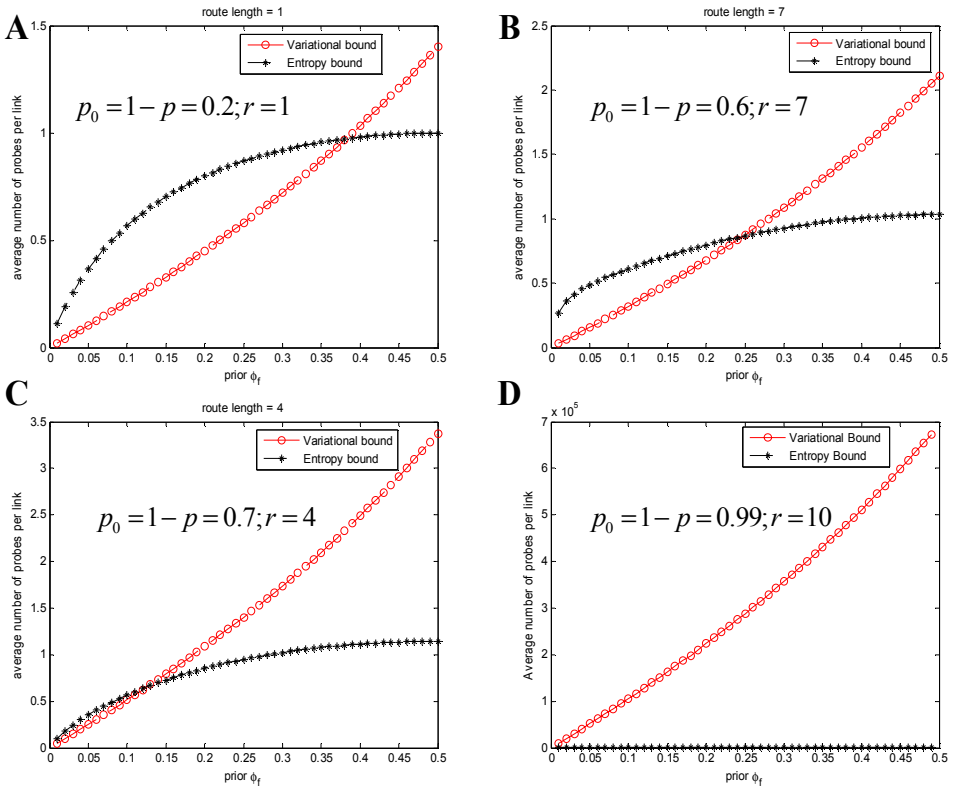


Figure 18. (A-D) Comparison of the variational and the entropy lower bound on the average number of probes per edge for varying noise values

Let the prior probability of link failure be ρ_f, $0 \le \rho_f \le 0.5$. Let ρ_1 be the probability that the probe succeeds in arriving at the ed-host even though a link on the path has failed, i.e., $\rho_1 = \prod_{j \in \pi_i} (1-p)^{1-x_j}; 1 \le i \le m_f$ and ρ_2 be the probability that the probe fails even though the links are functional due to measurement errors at the end-host and is equivalent to p_0. For a BSC, we have $\rho_{out} = (1-\bar{\rho})\rho_1 + \bar{\rho}(1-\rho_2)$, where $\bar{\rho} = (1-\rho_f)^r$ and r is the route length. In the absence of noise, we have $\rho_{out} = \bar{\rho}$.

Lemma 1: *For a network of n links, the information gained about the link state* $\boldsymbol{X}$ *by a single probe* Y *of length* r *in the presence of measurement and re-routing noise and the prior probability of link failure* ρ_f *is given by* $I(\boldsymbol{X};Y) \le H(\rho_{out})$, *where* $\rho_{out} = (1-\bar{\rho})\rho_1 + \bar{\rho}(1-\rho_2)$.

The proof is given in *Appendix E.* The above result provides information about the link status of the network, $\boldsymbol{X}$, in the presence of noise by measuring the output of a single probe Y of length r and thereby reduces the entropy of the network.

Lemma 2: *For a network of n_f links, the upper bound on the information gained about the link state $\boldsymbol{X}$ by m_f probes each of length $\{r_i\}_{i=1}^{m_f}$ is given by* $I(\boldsymbol{X};\boldsymbol{Y}) \leq \sum_{i=1}^{m_f} H((1-(1-\rho_f)^{r_i})(1-\rho_1)+(1-\rho_f)^{r_i}\rho_2)$.

The proof is given in *Appendix F.* The bound from *Lemma 2* provides information gained about the link status $\boldsymbol{X}$ by sending m_f probes of varying length and collecting the outputs at the end-hosts since our goal is to be able to obtain as much information about the link status using minimum number of probes (m_f). In the above result, as a special case, if all the probes have the same lengths, then $I(\boldsymbol{X};Y_1,Y_2,...Y_m) \leq m_f I(\boldsymbol{X};Y)$, where $I(\boldsymbol{X};Y) \leq H\lfloor(1-(1-\rho_f)^r)(1-\rho_1) + (1-\rho_f)^r\rho_2\rfloor$ is the information gained about the link status $\boldsymbol{X}$ using a single probe Y. In the absence of noisy observations it can be shown that $I(\boldsymbol{X};\boldsymbol{Y}) = m_f H(\bar{\rho})$.

Theorem 3: *For a noisy BSC, the lower bound on the average number of probes per edge (m_f/n_f) for asymptotic error free diagnosis is given as $m_f/n_f \geq H(\rho_f)/H(\rho_{out})$, where ρ_f is the prior probability of link failure and $\rho_{out} = (1-\bar{\rho})\rho_1 + \bar{\rho}(1-\rho_2)$.*

The proof is given in *Appendix G. Theorem 3* states that for a network with $|E| = n_f$ edges, if the number of observations is less than m_f, we are guaranteed not to achieve a zero diagnosis error asymptotically.

Next, the expression for optimal probe route length for the noisy case is given as (See *Appendix H*)

$$r^{opt} = \max_r \left\{ H\left(\bar{\rho}(1-\rho_2) + \bar{\rho}\rho_f/(1-\rho_f)\rho_1\right) - \bar{\rho}\left[H(\rho_2)\right] - \bar{\rho}\rho_f/(1-\rho_f)\left[H(1-\rho_1)\right]\right\}, \quad (40)$$

and is shown in Figure 19. If the prior probability of failure is small, the information gain can be maximized (entropy of the network can be minimized) by sending probes of longer lengths. As

the prior probability of failure or the noise parameters increase we have to revert to shorter length probes. In the noiseless case, (40) reduces to $r^{opt} = \max_r H((1-\rho_f)^r)$, which is compared with the result from [45] for varying prior in Figure 20, where the plot depicts an exact match.

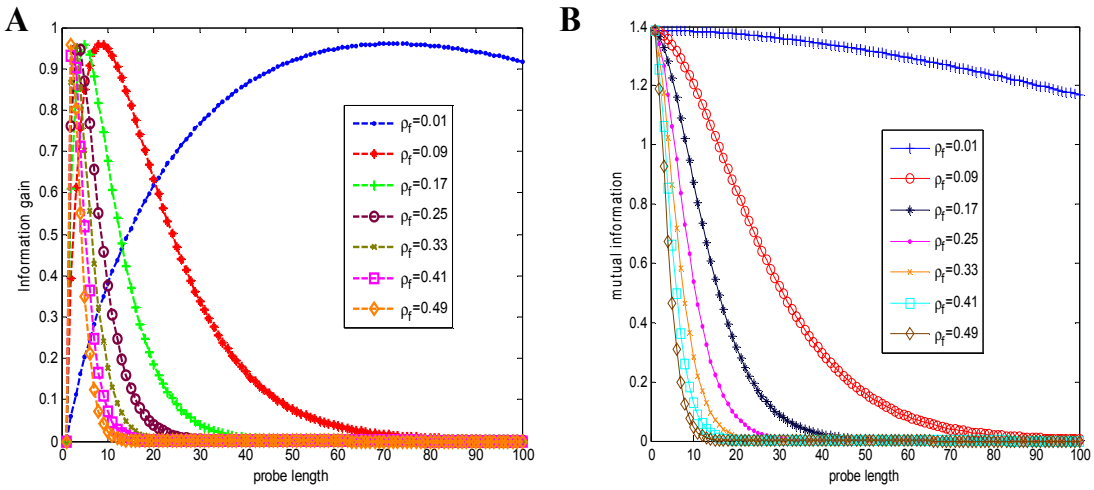


Figure 19. (A) Optimal probing length for varying prior ρ with noise ($1-p=p_0=0.01$), and (B) Optimal probing length for varying prior ρ with noise ($1-p=p_0=0.99$).

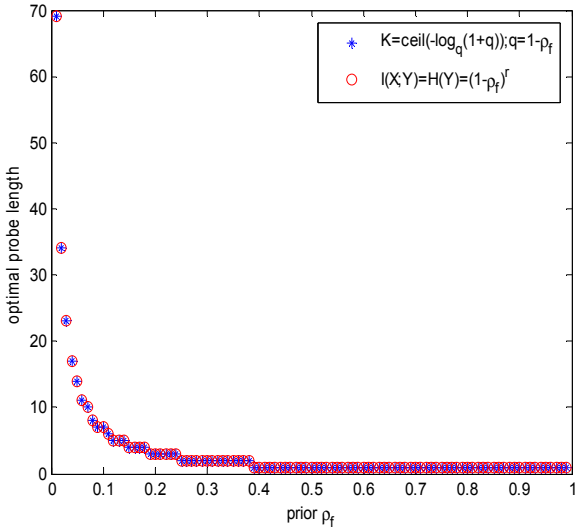


Figure 20. Optimal probe route length comparison

3.2.2 Numerical Results

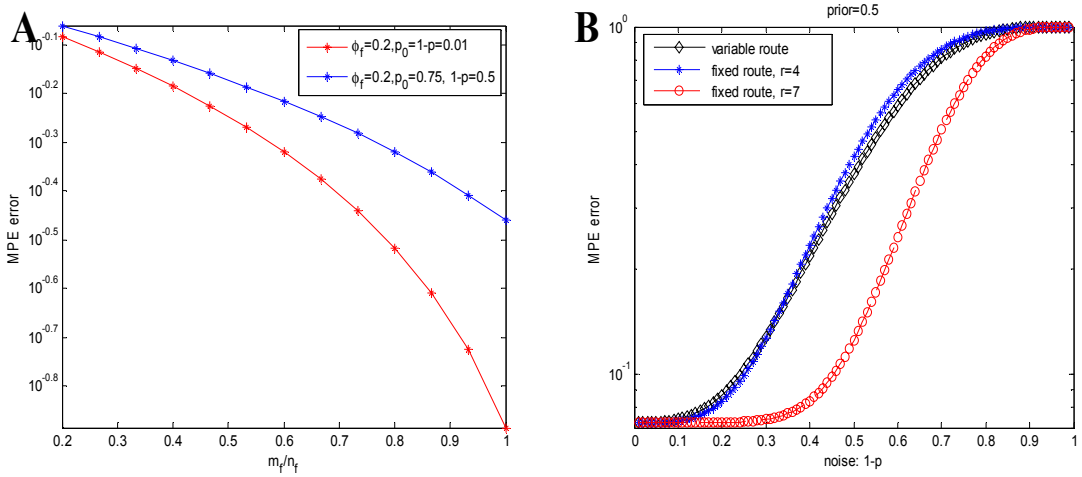


Figure 21. (A) Variational *MPE error* lower bound as a function of (m_f/n_f) and, (B) variational *MPE error* lower bound as a function of the re-routing noise

We validate the bounds using simulation for failure diagnosis. In particular, the simulation provides a comparison of the variational and entropy lower bounds on the average number of probes per edge and evaluates scalable/non-scalable using fixed/variable length probing. To assess the tightness of the variational lower bound, we generate a synthetic random topology of 15 nodes and 24 links using BRITE topology generator [82], where the routes were calculated using shortest path algorithm. We then employ greedy heuristic algorithm for obtaining the probing paths since this is well-known NP-hard problem [83]. The goal of this setup is to obtain a minimum number of paths that allows maximum network coverage. Figure 16 demonstrates that the bound is tight for large noise values, $p_0=1-p=0.99$ and moderate size networks. Our simulations were averaged over 1000 runs for the and in each run we randomly generated the link states and the *iid* noise parameters p_0 and $1-p$. Nonetheless, the result is insufficient to conclude about the tightness of the variational lower bound for large networks which is the limitation of this work. The difficulty is the exponential computational cost in evaluating the exact MPE error and BER for large networks for comparison. This is also the problem encountered in the related works [16, 45]. Nevertheless, the bounds provided for failure diagnosis are valid for any size network [84].

Table 1. Comparison of the constants for various lower bounds

VB Bound	*Constants*
MPE Error	$c_{MPE} = \left(\log\left[(1-p_0) + (1-p_0)^{-\xi}(1-p)^{-r\xi}\xi^{\xi}(1+\xi)^{-(1+\xi)} \right] \right)^{-1}; \delta_{f_1} = 0$
BER	$c_{BER} = \left(r_{\max} \log\left[(1-p_0) + (1-p_0)^{-\xi^{opt}}(1-p)^{-r\xi^{opt}}\xi^{opt\,\xi^{opt}}(1+\xi^{opt})^{-(1+\xi^{opt})} \right] \right)^{-1}; \delta_{f_2} = 0$
EL Bound	$c_{EL} = \left(H(\rho_{out}) \right)^{-1}; 1-\rho_{out} = (1-\bar{\rho})(1-\rho_1) + \bar{\rho}\rho_2 \;, \bar{\rho} = (1-\rho_f)^r$

Figure 21(A) depicts the decrease in the *MPE error* when the number of observations are increased as a function of the network size for a variable routing scheme with a prior $\phi_f = 0.2$ and noise values, p_0*=1-p=0.01* and p_0*=0.75, 1-p=0.5* respectively. The results were averaged over 100 independent runs. Figure 21(B) shows the effect of re-routing noise parameter *(1-p)* on the *MPE error* for fixed *(r=4 and 7)* and variable routing schemes. The MPE error decreases with the increase in probe route length and the number of probes since more information is gathered about the network state. In all the cases, we chose the number of observations (*m=24),* p_0*=0.01* and failure probability, $\phi_f = 0.5$. Figure 18 shows a comparison of the entropy lower bound from *Theorem 3* and variational MPE lower bound from (35) on the average number of probes per edge for a network size of n_f*=2000,* varying prior and noise values. A fixed routing scheme *(r=1, 4, 7 and 10)* was used for comparison. For the noiseless case, the bounds intersect at ρ_f *=0.5* and as the noise parameters increase the intersection point shifts to the left as seen from the figure. Hence, we can observe that the entropy bound is tighter for small failure probability and low noise values whereas the variational bound becomes tighter compared to the entropy bound as noise values increases. The comparison of the constants that determine the growth rate of the number of measurements as a function of the network size, route length and noise parameters (p_0 and *1-p*) are provided in Table 1. For the low-noise case, the growth rate of the measurements as a function of the network size is linear when compared to the non-scalable case in Figure 18(D).

However, if the prior probability of failure is known, then from *Theorem 3*, we would require $n_f\left(\frac{H(\rho_f)}{H(\rho_{out})}\right)$ observations asymptotically. But, if we do not have any information about ρ_f, then we have the following result.

Proposition 1: *The lower bound number of observations required to fully localize failed links if the prior probability* ρ_f *is unknown can be asymptotically given as*

$$m_f / n_f \geq \left(H(\rho_f) + \Omega_{n_f'}\right), \tag{41}$$

where $\Omega_{n_f'} = \min_{\hat{\rho}_f \in \Gamma_{n_f'}} \left(\frac{1}{n_f}, L_{n_f'}(\hat{\rho}_f) + D(\rho_f \,||\, \hat{\rho}_f)\right)$ *is the penalty term,* $L_{n_f'}$ *is the code length,* $\Gamma_{n_f'}$ is the space of probability distribution functions $\hat{\rho}_f(.)$ *and* $D(\rho_f \,||\, \hat{\rho}_f)$ *is the Kullback-Leibler distance.*

The proof is given in *Appendix I.* This result suggests that we would require on an average $n_f\left(\frac{H(\rho_f)}{H(\rho_{out})} + \Omega_{n_f'}\right)$ observations asymptotically to infer $\boldsymbol{X}$ (link status). Now the penalty term, $\Omega_{n_f'}$ which approximately provides the additional measurements that would be needed to infer the link status can be obtained following the result in [85, 86]. From [85], the average number of additional observations that would be required asymptotically is given by

$$\Omega_{n_f}(\rho_f) \sim \frac{1}{n_f}\left((\frac{\upsilon}{2})\log(n_f) + \log\int_{\Upsilon}\sqrt{\left|I(\rho_f)\right|}d\rho_f - \frac{\upsilon}{2}\log(2\pi) + o(1)\right), \tag{42}$$

where υ is dimension of the parametric family. The logarithm here is base 2. The term $\left((\frac{\upsilon}{2})\log(n_f) + \log\int_{\Upsilon}\sqrt{\left|I(\rho_f)\right|}d\rho_f - \frac{\upsilon}{2}\log(2\pi)\right)$ is known as the asymptotic parametric complexity (APC) in [85]. Solving the above for our case, we have $\upsilon = 1$, and the Fisher information associated with estimating ρ_f, $I(\rho_f) = \frac{1}{\rho_f(1-\rho_f)}$. Hence, $\log\int_{\Upsilon}\sqrt{\left|I(\rho_f)\right|}d\rho_f = \pi$. Therefore,

asymptotically, $\Omega_{n_f} \sim \frac{1}{n_f}\left(\frac{1}{2}\log(\frac{n_f\pi}{2}) + o(1)\right)$. A similar result is found in [86], which is termed as the index of resolvability and the expression is $\Omega_{n_f} \sim \frac{1}{n_f}\left(\frac{\upsilon}{2}\log(n_f)\right)$, which essentially is similar to the result obtained for $\upsilon = 1$.

To summarize, we have shown that variational technique is one of the methods to solve intractable Bayesian inference problems and provides us a tool to explicitly characterize the constant that multiplies the growth function. We would presume that the "sample complexity" results considered in this work should also be applicable to any other similar approximate inference techniques. Furthermore, the topology information is embedded in the bi-partite graph since the paths are chosen using a greedy heuristic such that it covers maximum number of links. The limitations of the diagnosis approach is that the exact MPE error and BER cannot be evaluated for large networks due to the exponential computational cost and also the simulation is currently limited to synthetic networks.

3.3 Congestion Diagnosis

We now consider congestion at nodes that can occur much more frequently than link failures. We assume that all nodes in the network are vulnerable to congestion, and hence each network node has two states, congested or not. Let us define the status of each node as X_i, i being the node index. Suppose $n_c = |V|$ then, $\boldsymbol{X} = \{X_1, X_2, ...X_{n_c}\}$. The status of each node can be defined as

$$X_i = \begin{cases} 1 & \textit{if not congested} \\ 0 & \textit{otherwise} \end{cases} . \tag{43}$$

Hence, we have the vector, $\boldsymbol{X} \in \{0,1\}^{n_c}$. In this problem, probe packets are assumed to be marked with the highest priority, and sent from different sources to destinations, where the destinations correspond to the edge-routers/end-hosts that are equipped with monitoring capability. One way to implement such a probing method is to send a probe packet that is a normal packet but has a unit bit-field that is initialized to zero and is flipped on its way through the selected probing paths

if a node is not congested. When the packet arrives at a congested node, because of its high priority it is just pushed to the beginning of a queue and is passed on to the next node without flipping. Once the packet arrives at the destination, its binary bit field is observed.

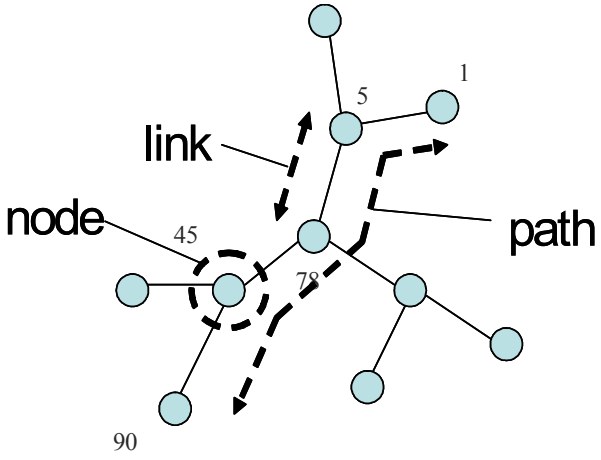


Figure 22. Graphical representation of a path

Consider Figure 15, where the shaded routers are congested. As before, we construct a bipartite graph as shown in Figure 15, where on one side we have the set of paths the probe takes and on the other side the nodes in those paths. The nodes in a particular path influence the output of that path and hence edges are directed from the nodes in the path to the output. The observation vector is the output of the probe packet bit field collected at the end-hosts. We point out that if n_c is the total number of nodes inside the given IP topology, the number of observations m_c needed is always less than or equal to n_c', where n_c' is the number of nodes on the measurement paths, where $n_c \geq n_c'$. For $m_c < n_c'$, the proposed approach can detect up to a certain fraction of congested nodes and the maximum fraction being when $m_c = O(n_c')$, ensuring zero error.

Consider the following illustrative example in Figure 22, where the packet path is defined by nodes $\{1,5,78,45,90\}$. Assume that in this path node 5 and node 78 are congested. We initialize the bit field in the probe packet to zero. Since nodes 1, 45 and 90 are not congested the bit field is flipped to one at node 1, and then at nodes 5 and 78 the bit remains unchanged and at node 45 the bit is flipped back to zero. Then at node 90 the bit is again flipped to one. So the

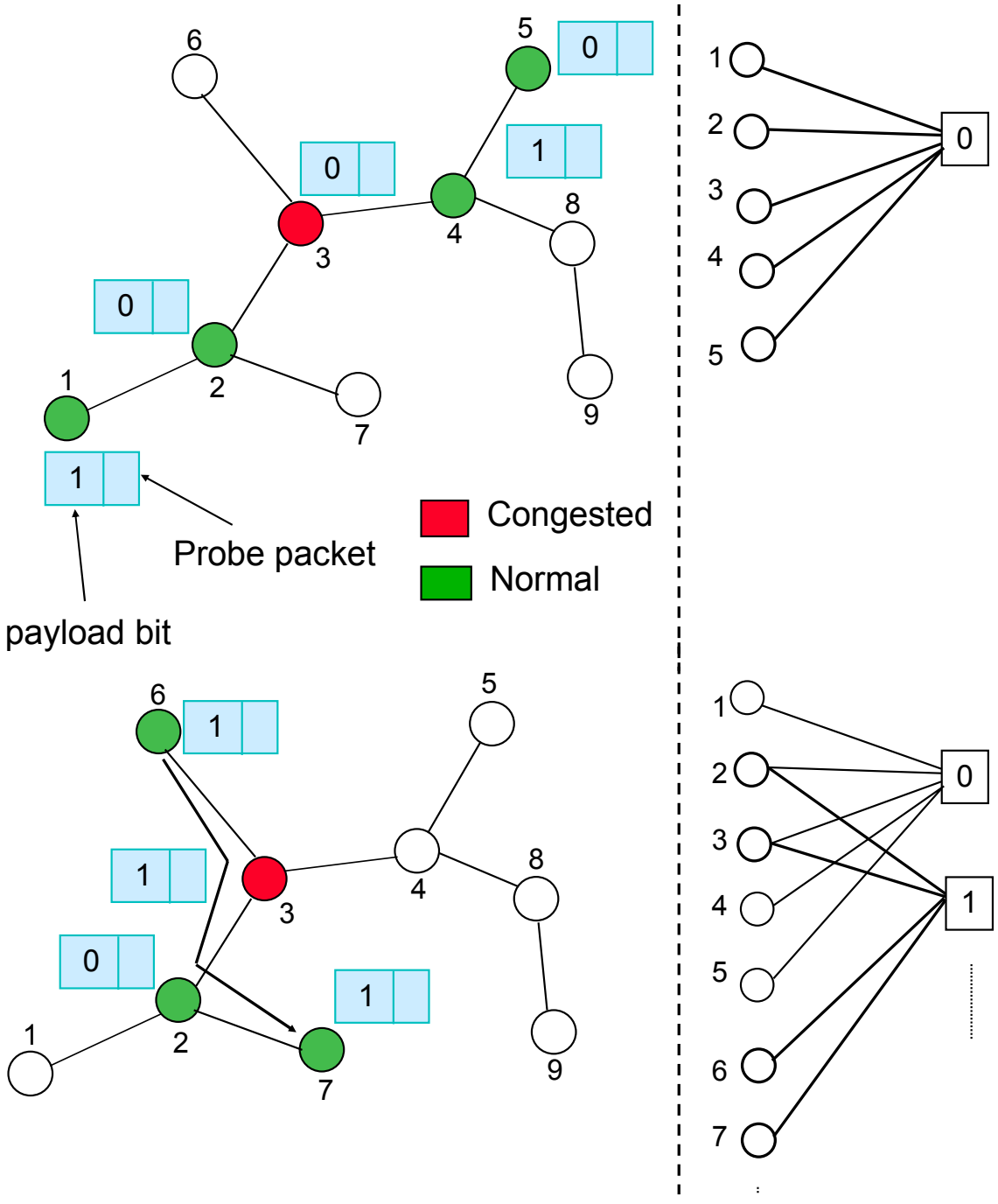


Figure 23. Representative example of creating a Tanner graph.

output of the bit field is $X_1 + X_5 + X_{78} + X_{45} + X_{90} = 1$, where '+' denotes the binary element-wise XOR operation, and is one of the observations in the vector y. The formation of bi-partite

(Tanner) graph is shown in Figure 23. Hence, the problem of state observation has been converted to solving a set of linear equations given a set of observations and is of the form

$$A\boldsymbol{x} = \boldsymbol{y}\,, \tag{44}$$

where A is the routing matrix of size $n_c \times m_c$ and has a '1' in the ij^{th} position if node i lies on path j. $\boldsymbol{y}$ is the set of observations measured at the edge routers, which is a column vector of size $m_c \times 1$ and $\boldsymbol{x}$ denotes the realizations of the random variable $\boldsymbol{X}$ and is of size $n_c \times 1$.

Usually the dimension of A is large, and in general it is not full-rank. Therefore, there are many possible solutions to equation (44) [18, 19], and we choose a maximum-likelihood solutions obtainable from (27) where, $\boldsymbol{x}$ has the maximum Hamming weight (maximum number of 1's). Moreover, $\boldsymbol{y}$'s are subjected to "noise" which arises because the routers that flip the bit in the probe packet may not function accurately and as a result the probe outcome is erroneous. The other reasons for measurement noise are due to route loops in the routing protocol, random unpredictable fluctuations in the competing traffic, and delay jitter in packet transmission [87].

3.3.1 Relation between the Congestion Diagnosis and ECC Framework

We now study the scalability issues of congestion localization, which can be considered as a soft network failure. Unlike in the case of failure diagnosis, the end-to-end probing now is based on a bit-flipping method and thus represents a logical XOR operation. Hence, we develop a new framework by mapping the congestion localization to problem of decoding error control coding (ECC) over a binary symmetric channel. We assume, (a) underlying IP topology is known. (b) each node gets congested independently of the other nodes. Now the network size is the total number of nodes rather than the links. As before, we treat the prior probability of congestion as a parameter in the range, $0 \le \rho_c \le 0.5$. Later, we investigate scalability issue in the presence of "noisy" measurements and when ρ_c is unknown.

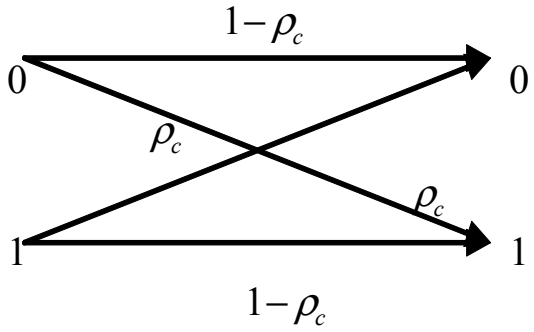


Figure 24. Binary Symmetric Channel model for congestion diagnosis

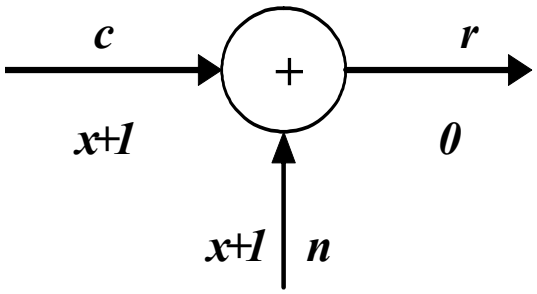


Figure 25. Channel model mapping to the congestion diagnosis problem

Let us consider a binary symmetric channel (BSC) as in Figure 24, with input symbols $X \in \{0,1\}^{n_c}$ and error probability ρ_c, which is also termed as crossover probability. The channel model is shown in Figure 25, where ***c*** is the codeword that is sent through the channel that adds noise vector ***n*** and the received codeword is ***r***. ***c*** is a *n*-bit codeword, which can be any one of the 2^k possible codewords, *k* being the message length. This is called the *(n, k)* binary code. For a *(n, k)* binary code the rate *R* is defined as *k/n*. From Figure 25, we have ***r*** = ***c***+***n,*** where '+' denotes the element wise binary XOR operation. Here, the noise vector ***n*** is independent with $\Pr\{n_i = 1\} = \rho_c; \Pr\{n_i = 0\} = 1 - \rho_c$. Let us consider a linear error correcting code, which has the following property

$$\Xi \boldsymbol{c} = 0, \tag{45}$$

where Ξ is the parity-check matrix of size $(n-k)\times n$ and has full row-rank. Given the received codeword $\boldsymbol{r}$ and the condition $\Xi\boldsymbol{c}=0$, we find the transmitted codeword $\boldsymbol{c}$ and the noise vector $\boldsymbol{n}$. Let $\Xi\boldsymbol{r}=\boldsymbol{d}$ and since $\boldsymbol{r}=\boldsymbol{c}+\boldsymbol{n}$, we have

$$\Xi\boldsymbol{n}=\boldsymbol{d}. \tag{46}$$

One possible method of obtaining the transmitted codeword $\boldsymbol{c}$ from $\boldsymbol{r}$ is by solving (46) by choosing the noise vector $\boldsymbol{n}$ with the minimum hamming weight.

3.3.2 Mapping the Congestion Diagnosis to the Error Correcting Code Problem

Consider equation (44), and assume that instead of $\boldsymbol{x}$, we are trying to find $\boldsymbol{x}^* = \boldsymbol{x}+\boldsymbol{1}$, where $\mathbf{1}$ is the all-one vector, i.e. we are trying to solve $A\boldsymbol{x}^* = \boldsymbol{y} + A\boldsymbol{1} = \boldsymbol{y}^*$. The vector of $\boldsymbol{1}$ is added to convert the congestion localization problem to the ECC problem. Therefore, as mentioned earlier, we have $\Pr\{X_i=0\}=\rho_c;\Pr\{X_i=1\}=1-\rho_c$, but if $\boldsymbol{x}^*=\boldsymbol{x}+\mathbf{1}$, then $\Pr\{X^*_i=1\}=\rho_c;\Pr\{X^*_i=0\}=1-\rho_c$. Let the matrix A and vector $\boldsymbol{y}^*$ be the same as the parity check matrix Ξ and vector $\boldsymbol{d}$ respectively as defined in (31). The elements of $\boldsymbol{x}^*$ are also independent and have the same distribution as noise $\boldsymbol{n}$ defined above. In the problem of decoding of ECC over BSC, as shown in Figure 25, suppose the transmitted word is $\boldsymbol{x}^*$ and the noise is also $\boldsymbol{x}^*$, the received word is all zero-vector $\mathbf{0}$. Note that we are dealing with non-linear ECC, since $\Xi\boldsymbol{x}^*(A\boldsymbol{x}^*)\neq 0$. Hence, if we could solve $A\boldsymbol{x}^* = \boldsymbol{y}^*$, we can also solve $\Xi\boldsymbol{n}=\boldsymbol{d}$, and be able to decode the transmitted codeword $\boldsymbol{c}$. The information capacity of a binary symmetric channel with bit-error probability ρ_c is [88]

$$C=1-H(\rho_c). \tag{47}$$

Shannon's second channel coding theorem implies that [88], unless the rate $R\le C$, it is not possible to determine the codeword $\boldsymbol{c}$ from $\boldsymbol{r}$ with probability of error tending to zero asymptotically (i.e. when the length of $\boldsymbol{c}$, $\boldsymbol{r}$ tends to infinity). From (47), we have

$$1-R\ge H(\rho_c). \tag{48}$$

For a parity-check matrix Ξ (i.e., A in the congestion localization problem), 1-R is the ratio of number of rows to the number of columns in Ξ where R is the rate of the code. Thus, (48) gives a

lower bound on the ratio of the number of independent observations and the number of network nodes to determine the state of every network node with zero probability of error (asymptotically). The results of the following theorems are applicable in the congestion localization problem. From the result in [89] (cf. Theorem 5), we have

$$1-R \geq \frac{1-C}{\sum_d p_d H(\rho_d)}, \tag{49}$$

where $\rho_d = \frac{1}{2}\left(1-(1-2\rho_c)^d\right)$. Now, using equation (47) the above equation reduces to

$$1-R \geq \frac{H(\rho_c)}{\sum_d p_d H(\rho_d)}, \tag{50}$$

where $p_d = \frac{w}{n_c'}$ is the degree distribution on the variable side. Here, *w* is the number of nodes with degree *d* and n_c' is the total number of nodes on the variable side. In the congestion localization problem, the minimum set of probing paths forms the check side and the network nodes in these paths form the variable side, as shown in Figure 15. In our case, $n_c' \leq n_c$, where n_c denotes the total nodes in the physical IP topology. From the bipartite graph construction, we have

$$1-R = \frac{\#\ of\ observations}{\#\ of\ network\ nodes} = \frac{m_c}{n_c'} \geq \frac{H(\rho_c)}{\sum_d p_d H(\rho_d)}. \tag{51}$$

Equation (51) implies

$$m_c \geq c_c n_c', \tag{52}$$

where $c_c = \frac{H(\rho_c)}{\sum_d p_d H(\rho_d)}$ is a constant and $0 \leq c_c \leq 1$. That is, the number of measurements needed to identify all congested nodes on the bipartite graph grows linearly with the network size and we require at least m_c probe packets are needed to fully identify the nodal status on the bipartite graph. Given the ratio $\frac{m_c}{n_c'} = 1-R = 0.5$, the maximum fraction of congested nodes (ρ_c *) that can be localized without error is *11%* from equation (48). Since the routing matrix *A*

does not have equal row weights, the ρ_c that can provide zero detection error is less than ρ_c*. We use the term detection error in this case since detection and localization is performed together. Hence if we fix the set of observations (probing paths), the probability density function of the check side, p_d, and the number of nodes, n_c' are known, and hence ρ_c can be obtained. On the other hand, varying the percentage of congested nodes, ρ_c then the growth rate of the number of required measurements can be determined with respect to the network size that can result in zero detection error. Therefore, the bound in (52) provides us the measure of scalability. When the code length is finite as opposed to (48), which is valid asymptotically, the sphere-packing lower bound for a *(n, k)* code on the BSC with the cross over probability ρ_c holds, [90] (Page 163)

$$P(n,k,\rho_c) = \sum_{i=r+1}^{n} \binom{n}{i} \rho_c{}^{i} (1-\rho_c)^{n-i} - \rho_c{}^{r+1} (1-\rho_c)^{n-r-1} \left(2^{n-k} - \sum_{i=0}^{r} \binom{n}{i} \right), \tag{53}$$

where *r* is defined as the maximum integer such that $\sum_{i=0}^{r} \binom{n}{i} \le 2^{n-k}$ and $P(n,k,\rho_c)$ is the pattern error rate which is defined as the ratio of number patterns in error (single/multiple bit errors) to the total number of patterns. Equation (53) provides a lower-bound for finite length codes where the rate $R=k/n$ and $q = \frac{\#\ of\ observations}{\#\ of\ network\ nodes} = 1 - k/n$.

3.3.3 Congestion Diagnosis in the Presence of Noise and Unknown Prior

We now consider noisy probe observations where the sources of noise in the case of congestion diagnosis are explained earlier. Considering the results from *Theorem 3*, for a noisy BSC, we have $m_c / n_c' \ge H(\rho_c)/H(\rho_{out})$, where n_c' is the number of nodes on the probing paths. Hence, we have, $\rho_{out} = \rho_1 + \bar{\rho}(1-2\rho_1)$. From (51), we can obtain the following relationship

$$1 - R = \frac{m_c}{n_c'} \ge H(\rho_c)/H(\rho_{out}). \tag{54}$$

Using the expression for optimum probe route length from (40) we can obtain the maximum fraction of congested nodes (ρ_c*) that can be localized without error. In the noiseless case, $H(\rho_{out}) = H(\bar{\rho}) = H(1-\rho_c)^r$ and using optimal probe route length expression

$r^{opt} = \max_r H((1-\rho_c)^r)$, $H(\rho_{out}) = 1$ and hence we arrive at equation (48). In the case of unknown prior, the result from *proposition 2* holds. If we take the graph structure into consideration, then we have $\frac{m_c}{n_c'} \geq \left(\frac{H(\rho_c)}{\sum_d p_d H(\rho_d)} + \Omega_{n_c'} \right)$. Hence the penalty term $\Omega_{n_c'}$ indicates an additional fraction of measurements needed to account for unknown probability of congestion, thereby reducing the percentage of congested nodes that can be localized with zero error as compared to (51).

3.3.4 Simulation Results

For the congestion diagnosis case, we consider three different probing experiments namely noiseless, noisy and when ρ_c is unknown and compare the results against the fundamental limits. To identify the congested nodes we construct a factor graph, and congestion is inferred using the message-passing algorithm which is computationally efficient [91, 92]. Hence, the simulation validates the theory and provides insights into practical limitations that arise from the network size, the graph structure, the noise parameters and the prior probability of congestion.

3.3.4.1 Congestion Diagnosis on Large Networks using Belief Propagation Algorithm

We use a belief propagation (BP) algorithm (also called message-passing algorithm) [48] to obtain the status of the network nodes. The message-passing algorithm approximates the *maximum a posteriori* probability (MAP) by constructing a bipartite graph that consists of the variable and the check side as shown in Figure 15. The localization algorithm can be given as follows:

1) Generate the topology *G=(V, E)*.

2) Find the shortest paths $\forall v_i \in V$ to $v_j; j \neq i$,

3) Select the paths of G such that maximum nodes $|V'|$ are covered such that $|V'| \leq |V|$.

4) Using (51), compute the maximum percentage of congested nodes that can be localized

with zero detection error.

In the context of decoding error-control codes, the BP algorithm is an iterative algorithm that works on bipartite graphs to compute (approximately) the maximum-likelihood (ML) estimate of the transmitted codeword. For Low Density Parity Check (LDPC) codes, the graph shown in Figure 15 is called the Tanner graph. Tanner graphs are constructed from the parity check matrices of the corresponding LDPC codes. The rows in the parity check matrix are denoted by the nodes on one side (check nodes), and the columns are denoted by the nodes on the other side (variable nodes) of the Tanner graph. An edge exists if the corresponding entry in the parity check matrix in a 'one' Parity check matrices of LDPC codes are sparse matrices. Hence, the corresponding Tanner graphs are also sparse. The sparseness of Tanner graphs and the minimum number of cycles are the reason the message-passing algorithm approximates the ML algorithm closely, and are computationally much less intensive. In the LDPC decoder, log-likelihood ratios (LLR) are computed from the channel outputs first. These are the inputs to the message-passing decoder. The decoder typically runs several to dozens of iterations where messages (extrinsic LLRs) travel on the edges of the Tanner graph and are combined at the variable and the check nodes.

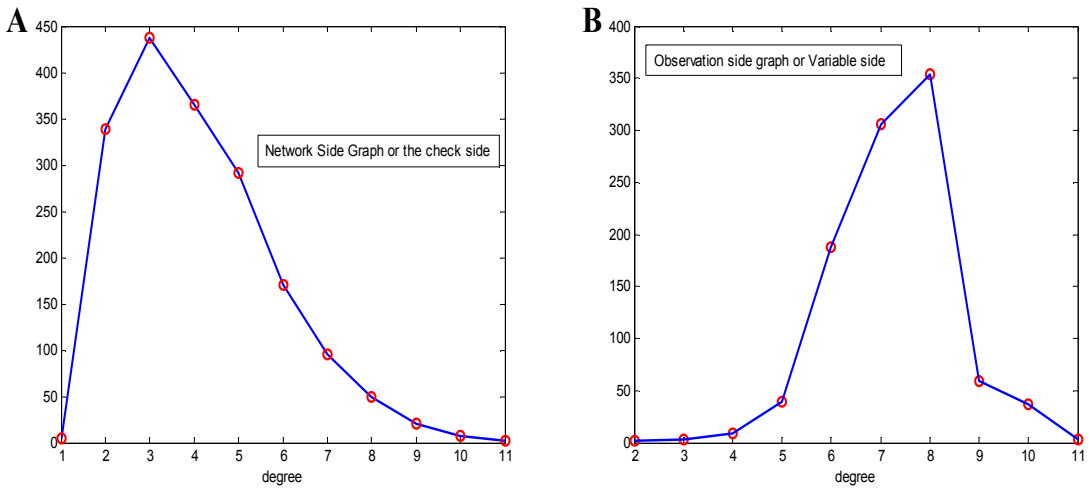


Figure 26. Degree distribution on the (A) check side (B) variable side for a network of size 1783 nodes and the number of observations=1000

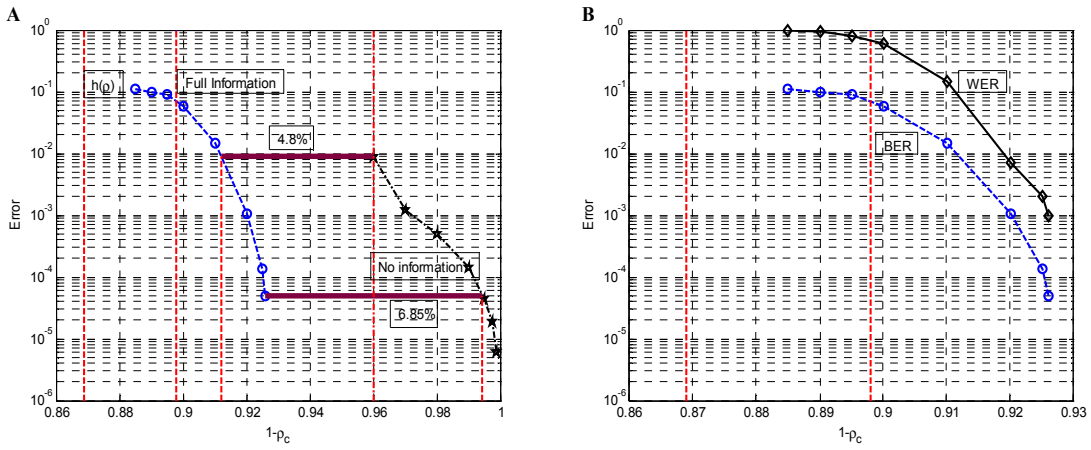


Figure 27. A) Detection Error vs. prior (ρ_c) with and without the knowledge of the prior probability of congestion and (B) BER vs. WER comparison for 1000 observations.

Simulations were performed on a random graph generated using BRITE topology generator [82] of various sizes, i.e., n_c=500, 1000, 2000 and 5000. This is similar to increasing the code length in the ECC framework. The paths were chosen according to shortest path routing algorithm from every node to every other node. Once the paths are obtained, these paths are ordered and selected in such that maximum number of nodes is covered. We employ a greedy heuristic algorithm for obtaining the probing paths as before. Once the paths have been chosen, we also need to reduce the number of cycles in the Tanner graph so that the message-passing algorithm works efficiently. In the optimal case, the message-passing algorithm assumes that the bipartite graph is cycle-free, which is not possible in this setting. Due to greedy heuristics and graph pruning that reduces the number of cycles; we cannot have all the nodes on the selected probing paths. Hence, the topology information is embedded in the factor graph. Furthermore, the paths are chosen such that observation side degree distribution corresponds to the hop count distribution observed on the Internet [81].

Our congestion localization algorithm is closely related to the message passing decoder for LDPC codes. Like Tanner graphs of LDPC codes, our bipartite graphs are reasonably sparse. The observations collected by the probe packets are used to combine LLR messages at the check nodes (observation nodes in our case). Unlike LDPC codes, we do not have any channel outputs to initialize the congestion localization algorithm. Hence, we initialize all variable nodes (network

nodes in our case) in the graph with the same value. The LLR messages are combined differently at the observation nodes depending on whether the corresponding observation is a 'zero' or a 'one' In the case of noisy measurements, we modify the localization algorithm to include the noise in observations. The "noise" here refers to the error (binary inversion) in the probe measurement at the end host. Unlike in the noiseless case, observation values now come with a certain probability (depending on the length of the probe and the noise probability) of being correct. Hence, the LLRs at the observation nodes are computed separately assuming binary possibilities, and then are weighted appropriately depending on the observation noise probability for that observation node. For the case when the prior probability of the network nodes being congested is unknown, the problem is similar to the decoding error control codes over binary symmetric channels with unknown error probabilities. The messages traveling on the edges of our bipartite graph are no longer LLRs but 'hard' values and the algorithm then iteratively combines messages in the bipartite graph. The performance of the algorithm is, of course, worse than the case where the prior probabilities are known.

In particular, simulations were averaged over 1000 patterns and the detection error was computed. Using greedy heuristics and 1000 observations we could cover 1783 nodes (89.2%) out of the total 2000 nodes. Hence, $1-R=0.561$, and from the bound in (48), we can guarantee that 13.1% (234 randomly chosen congested nodes out of 1783 nodes on the bipartite graph at each iteration) of the congested nodes can be localized with zero detection error provided the network is large and the H matrix is well designed. The network side (check side) and the observation side (variable side) degree distribution are shown in Figure 26(A) and (B) respectively. From equation (51) the bound is reduced to 10.2% (182 randomly chosen congested nodes out of 1783 nodes on the bipartite graph at each iteration) because of the scaling factor $\sum_d p_d H(\rho_d)$. This implies that below the lower bound ($\rho_c^* = 10.2\%$), we can guarantee that all congested nodes can be localized with zero detection error. From simulation we could localize all errors up to $\rho_c = 7.4\%$ and the results are shown in Figure 27(A), where $\rho_c^* = 10.2\%$ and the detection error shows a steep fall after 7.4% . Furthermore, if we could tolerate a finite detection error, δ_c, then a larger fraction of congested nodes (ρ_c), can be localized. For e.g., in

Figure 27(A), say if $\delta_c = 10^{-3}$, then we can localize up to 8% instead of 7.4%. The BER vs. WER comparison is shown in Figure 27(B) which depicts a similar fall-off. Scalability results are shown in Figure 28(A) where the y-axis represents m_c that are required to detect the congested nodes and the x-axis represents ρ_c that ensures zero detection error.

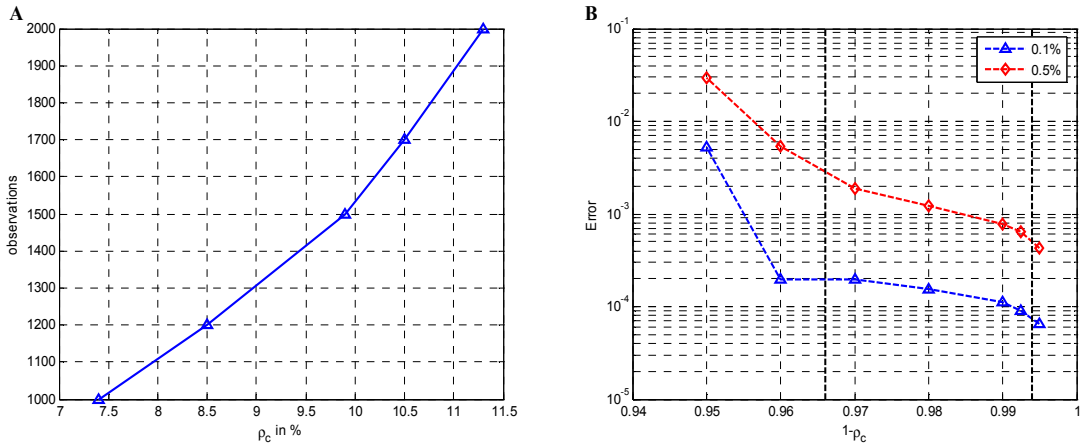


Figure 28. (A) Scalability Analysis, and (B) Detection Error vs. prior (ρ_c) with noisy observations for a network of size 1783 nodes and the number of observations=1000.

For the case of noisy probe observations, the results are shown in Figure 28(B) for noise prior of 0.1% and 0.5% respectively. It can be seen that unlike the noiseless case, the error decreases slowly with increasing prior. We believe that longer probe route lengths are responsible for this phenomenon since they have a higher probability of affected by noise. The theoretical bound from (54) for $1-R=0.561$ (i.e., with 1000 measurements), and ρ_1 *=0.1%* guarantees that 10.1% of the congested nodes can be localized with zero detection error. From simulation we could localize errors up to ρ_c = 1% and if we set $\delta_c = 10^{-3}$, then we can localize upto ρ_c = 4.5%. For $1-R=0.561$, and large noise probability, ρ_1 *=0.99%,* the bound in (54) guarantees that 0.1% of the congested nodes can be localized with zero error. Hence, we need much more than 1000 observations to increase the percentage of localization. In the case when the prior probability of congestion is unknown, the simulation results are shown in Figure 27(A) where it can be observed that only ρ_c = 0.8% could be localized with zero error. In the figure, we have also

compared the gain in terms having the full information as opposed to no information about the prior probability of congestion where the maximum deviation is about 6.85%.

To summarize this section, we show that the number of measurements $m_c \geq c_c n_c'$, where c_c is a constant which is explicitly characterized by relating the congestion localization in networks using a bit-flipping mechanism to the problem of decoding linear error correcting codes (ECC) over a binary symmetric channel (BSC) in coding theory. We also show that the growth rate of the number of measurements is of $O(n_c')$ and $0 \leq c_c \leq 1$ when the congestion probability is $\rho_c < 11.3\%$ for noiseless case. The growth rate is also of $O(n_c')$ for small noise parameters and when the noise parameters increases, the constant, $c_c > 1$. When the congestion probability is unknown, we have $c_c \gg 1$. The simulation results deviate from the theoretical bound for the following reasons. The result in (51) is asymptotic, i.e., when the code length (i.e., network size) tends to infinity. When the code length is finite, the sphere-packing lower bound for a *(n, k)* code on the BSC with the cross over probability ρ_c holds [90]. Also, if the graph has cycles in the case of congestion localization, the message passing algorithm no longer gives the exact MAP result, but it is an approximate solution and hence the performance deviates from the theoretical bound. Therefore, from the graph in Figure 27(A) we can observe that we are about 2.8% (10.2%-7.4%) away from the bound.

3.3.4.2 Computation and Implementation Cost

The computational complexity of the message-passing algorithm is proportional to $O(n \log n)$ where *n* is the block length of the code. In the congestion localization problem, the computational complexity depends on the number of network nodes in the bipartite graph. For 1000 iterations on a random graph of size 2000 and 1000 observations, our algorithm took roughly about 30 seconds on a Intel P4 machine with 1GB RAM. Measurements were performed using probe packets that have a payload of one bit. Hence the size of each probe packet is 41 bytes including the header. TCP traffic consists of packets of three sizes: 40 byte packets (the minimum packet size for TCP) that carry TCP acknowledgments without any payload; 1500 byte packets which is the maximum Ethernet payload size from TCP implementations that use path

Maximum Transmission Unit (MTU) discovery and 552 byte/576 byte packets from TCP implementations that do not use path MTU discovery. Hence, even if every packet were considered as a probe, the additional bit payload accounts for 2.73% of the maximum packet size. If probe packets were designated for network monitoring only, the traffic volume arising from these packets would be much smaller than the total flows arising from actual traffic and hence does not burden the network. This approach is practically realizable on networks where priority queuing algorithms are implemented. The two main differences between the current approach and Explicit Congestion Notification (ECN) [93, 94] are that: (a) Our approach flips the bit only when the router is not congested thus avoiding burdening core routers during congestion and, (b) In the ECN-setup if one or more routers are congested the bit is always set to 11, but in our case, due to the bit-flipping, we are able to localize one or more congested routers precisely using just a single bit in the probe packet using end-end measurements. We also use random reordering of the paths before we choose the set of paths to monitor and since each path is equiprobable; the measurement load on each end-host is similar or rather evenly distributed.

3.4 Summary

We have defined and investigated the scalability of end-to-end measurement-based network monitoring in the context of failure and congestion diagnosis [95]. We have developed a framework based on machine learning and information theory to study these problems. For failure diagnosis, a lower bound on the average number of probes for asymptotic error free failure diagnosis under noisy measurements is provided using variational methods. The analysis is validated by simulation results. Motivated from source coding, an entropy lower bound was derived and compared against the variational lower bound in the presence of noise. The constant that multiplies the growth function is derived explicitly and shown to depend on the failure probability, the measurement noise probability and re-routing noise probability. Next, we have formulated congestion localization as decoding of linear error control codes over a BSC and provided theoretical lower bounds. Simulations were performed on network of various sizes and under varying probing scenarios to verify the scalability result. The inference is performed using message-passing that has a computational complexity of $O(n\log n)$ (n being the code length) by

constructing a factor graph from the probed paths and the nodes on those paths. The sample complexity is shown to grow linearly with respect to the network size in both fault and congestion diagnosis under certain conditions. In the current approach for congestion localization, only one additional bit of payload is needed in the probe packet and hence the traffic volume generated due to probe packets does not burden the network.

CHAPTER 4

QUANTITATIVE EVALUATION OF DENOISING ALGORITHMS FOR BIOLOGICAL ELECTRON TOMOGRPAHY

This part of the dissertation investigates automatic techniques for quantitative analysis of biological specimens imaged using electron tomography. Our main focus is to apply information theoretic and learning approaches in mining relevant byte-sized information from gigabyte-sized tomograms. Specifically, this chapter we quantitatively evaluates and identifies robust denoising methods for biological tomograms that function effectively at low signal-to-noise ratios to maximize retrieval of information. Using tomograms collected from stained specimens at room temperature and from frozen hydrated specimens at cryogenic temperatures, in Section 4.1, we perform a spectrum of denoising algorithms based on nonlinear anisotropic methods, wavelet-based methods, and filtering-based techniques, without making *a priori* assumptions about the statistics or the type of noise present. We then identify the optimal denoising strategy for each tomogram using quantitative measures in Section 4.2, using goodness-of-fit (GOF) test based on Kullback-Leibler (KL) distance, Fourier ring correlation, and single-image SNR. In Section 4.3, the value of denoising for tomogram interpretation is assessed by qualitative and quantitative comparison of automated feature extraction versus the information retrieved manually by an experienced user from the same tomograms.

Previous studies using nonlinear anisotropic methods [96-98], wavelet based methods [99-101], and filtering [102] have already demonstrated the value of image denoising in various 2D and 3D datasets. A correlation-based automatic particle detection approach is investigated in [72], where anisotropic diffusion is used to enhance the edges and overall shape of the particles while reducing the noise. An improvement of the nonlinear anisotropic diffusion method for denoising in terms of reliable estimation of the derivatives and reduction in computation time and memory requirements is presented in [103], while Moss et al. [104] suggest using a 3D wavelet-based filter for preferentially highlighting objects of defined size-classes within 3D volumes. A dynamic bilateral filter for denoising low SNR images corrupted with impulse noise, while preserving edges by the use of an additional photometric exclusion function has been investigated

by [105] for cryo-electron microscopic images. This approach was an improvement over the bilateral filter that was proposed by Jiang et al.[106]. Development of useful denoising methods usually involves validating them in test cases where the statistics of noise is known [102, 107]. However, while the use of some type of denoising often qualitatively improves image quality, it is not obvious how to choose an appropriate denoising algorithm for a given biological tomogram to maximize feature visualization and extraction of quantitative information. The data quality of such tomograms can vary widely from differences in specimen preparation, data collection schemes, or origins of contrast as well as from using different algorithms to align and reconstruct the data into a final 3D volume. Given that any particular denoising algorithm may not perform uniformly well on such diverse datasets, it is essential to filter tomographic data based on analysis of the noise in a given dataset rather than *a priori* assumption of a particular noise model.

4.1 Denoising Methods

A variety of denoising methods were tested as briefly summarized below. We first compared the performance of non-linear anisotropic diffusion (NAD) [96, 107] and complex diffusion algorithms [98] on electron tomograms as indicated below. The diffusion equations in 2D are given as,

$$\nabla I = \nabla(G_\sigma * I(x,y,t)), \tag{55}$$

where G_σ is a Gaussian with standard deviation σ (regularization parameter).The equations for non-linear and non-homgenous diffusion [96] in 2D are given as

$$\frac{\partial I}{\partial t} = \nabla.(c(x,y,t)\nabla I), \tag{56}$$

where,

$$c(x,y,t) = 1/\left(1 + \frac{|\nabla I|^2}{\kappa^2}\right) \tag{57}$$

or

$$c(x,y,t) = e^{-\frac{|\nabla I|^2}{2\kappa^2}}. \tag{58}$$

Here, κ is a conductance parameter and $I = I(x, y, t)$ is the 2D tomogram. In the case of Weickert's equation [107], the gradient in c is a positive semi-definite diffusion tensor, varying with both magnitude and direction of the gradient. The diffusivity can be written in terns of the eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, and eigenvalues λ_1 and λ_2 as follows,

$$c = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1^T \\ \boldsymbol{v}_2^T \end{bmatrix}. \tag{59}$$

The eigenvectors are defined as $\boldsymbol{v}_1 \| \nabla I$ and $\boldsymbol{v}_1 \perp \nabla I$, and $\lambda_2 = \psi; 0 \le \psi \le 1$, to allow smoothing in the $\boldsymbol{v}_2$ direction and λ_1 can be chosen according to,

$$c(x, y, t) = \begin{cases} 1 & if \ |\nabla I| = 0 \\ 1 - \exp\left(-\dfrac{C_m}{\left(|\nabla I|^2 / \kappa^2 \right)^m} \right) & if \ |\nabla I| > 0 \end{cases}. \tag{60}$$

For the complex diffusion, we used

$$\frac{\partial I}{\partial t} = \nabla . (c(\mathrm{Im}(I)) \nabla I) \tag{61}$$

with

$$c(\mathrm{Im}(I)) = e^{i\theta} / \left(1 + \left(\mathrm{Im}(I) / k\theta \right)^2 \right), \tag{62}$$

where $\mathrm{Im}(I)$ is the imaginary value of I, θ is the phase angle, and k is the threshold parameter. The unprocessed tomogram is used as the initial condition for the above non-linear partial differential equations. Complex diffusion controlled by the signal's imaginary value avoids the stair casing effect that is characteristic of gradient-controlled nonlinear processes such as the Perona-Malik process [98]. Also, as pointed out by Gilboa *et al*., the real function is effectively decoupled from the imaginary one, and behaves like a real linear diffusion process, whereas the imaginary part approximates a smoothed second derivative scaled by time, and serves as an edge detector when the complex diffusion coefficient approaches the real axis. Complex diffusion also enables better performance in different nonlinear tasks such as ramp denoising.

In another well-known method of denoising, the phase data is preserved [100] by applying the continuous wavelet transform and using log Gabor functions to construct symmetric/anti-symmetric wavelet filters. The use of these functions permits the construction of

large bandwidth filters while still preserving a zero DC component in the even-symmetric filter which thereby minimizes the spatial spread of wavelet response to signal features. This leads to energy compaction in a limited number of coefficients. The process then determines a noise threshold at each scale, and by shrinking the magnitudes of the filter response vectors, phase information can be preserved after reconstruction.

Additional denoising methods we tested are based on translation-invariant wavelets with soft thresholding [99], Bayes least-squares Gaussian scale mixture method (BLS-GSM) [101], median filtering, and Wiener filtering [102]. To overcome the translation invariance problem of the wavelet basis, which causes visual artifacts (as in the case of Gibbs phenomena in the neighborhood of discontinuities), we used the "cycle spinning" method [99], where the translation dependence is averaged out by applying a range of shifts to the signal and thresholding the wavelet transform coefficients belonging to all the shifted versions of the signal. This can be applied as follows:

$$\bar{T}(\boldsymbol{x}) = Ave\left[S_{\Delta}^{-1}(T(S_{\Delta}(\boldsymbol{x}))\right], \tag{63}$$

where $\boldsymbol{x}$ denotes the multidimensional noisy data, T is the denoising operator, S_{Δ} is the translational shift operator, and Δ is the translational degrees of freedom in the signal space. The time complexity of such an operation is $O(n\log_2(n))$ for n circular shifts. The Gaussian scale mixture (GSM) also takes translation invariance problem into account. Best results were obtained with the original steerable pyramid [101], which uses 8 orientations in a set of scales, plus 8 oriented high-pass subbands. Neighboring coefficients are then clustered which is possible as a result of multiplication by a scalar multiplier, due to which the wavelet coefficients of image features are correspondingly amplified or damped. Clustering of the coefficients is obtained using the Gaussian-scale mixture model as follows:

$$\boldsymbol{y} = \sqrt{z}\boldsymbol{u}, \tag{64}$$

where $\boldsymbol{y}$ represents a local cluster of N wavelet neighbor coefficients arranged in a vector, $\boldsymbol{u}$ is a zero-mean Gaussian vector of given covariance and z is a hidden, independent, scalar random variable controlling the magnitude of the local response $\boldsymbol{y}$. The random vector $\boldsymbol{y}$ is termed a Gaussian Scale Mixture (GSM). Then for every chosen z and noise vector $\boldsymbol{r}$, we calculate the

value of the posteriors, $p(z \mid r)$, using Bayes rule since $p(r \mid z)$ and $p(z)$ are known. It is assumed that the individual noise vectors are independent and additive Gaussian samples. This strategy provides a smaller quadratic error than the classical (empirical Bayes) approach, which first estimates the hidden variable, and then applies a least-squares estimator assuming the estimated value was exact [101]. We adapted the publicly available Wavelab [108] and BLS-GSM MATLAB[109] [109] software to implement both methods mentioned above with minor modifications.

4.2 Quantitative Measures of Denoising

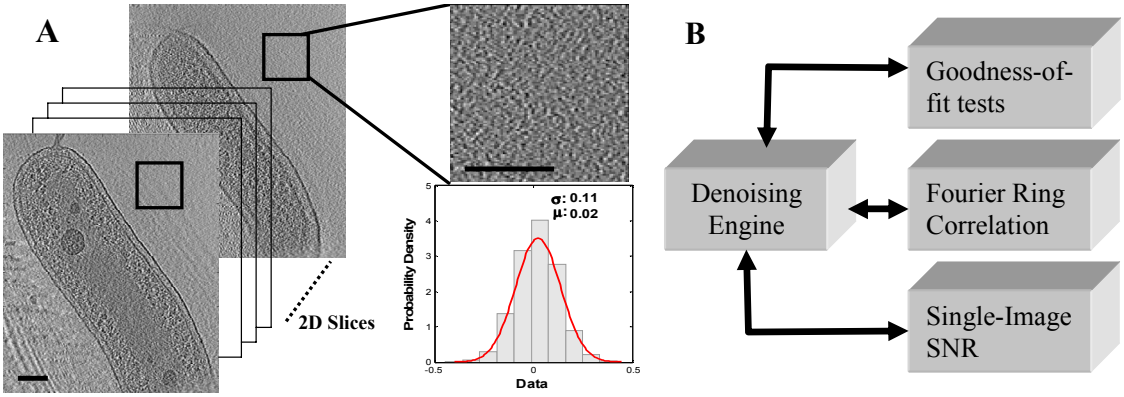


Figure 29. Strategy to evaluate algorithms used to denoise a noisy tomogram Y containing n 2D slices (A), i.e., $Y = (Y_i)_{i=1}^{n}$, where the index i represents the i^{th} 2D slice in the noisy tomogram. $\hat{Y}$ is the subset of Y, which is selected from a comparable region of each tomographic slice that contains no cells and hence no biological data (black box, enlarged on right). N represents the tomographic noise ($N = Y - X$, where X denotes the denoised tomogram). $\hat{N}$ is the subset of N similar to $\hat{Y}$, where no biological data exists. A Gaussian fit computed with the shown mean and variance to the noise samples in $\hat{Y}$ is shown on lower right (representative slice). Scale bar: 100 nm (B) Iterative process to identify the best denoising algorithm with optimized parameters, for a particular tomogram. Quantitative assessment was performed using analysis such as the KL-distance based GOF test, Fourier Ring Correlation and Single-Image SNR.

Previous studies using nonlinear anisotropic methods, wavelet based methods, and filtering based techniques, have already demonstrated the value of image denoising in various 2D and 3D datasets as outlined earlier. However, most of these algorithms are condition-specific, application-oriented, and assumption-based. A single denoising algorithm may therefore not perform uniformly well on diverse datasets that have been collected using a variety of specimens

and acquisition conditions, because of variations in noise patterns. Variations may occur from the use of low doses, errors in image alignment and reconstruction, and the anisotropic resolution due to the missing wedge problem in electron tomography.

In electron tomography, we do not have an *a priori* knowledge of the type of noise in the tomogram or its statistics. As explained below, we establish quantitative measures such as Kullback-Leibler (KL) distance [27], based goodness-of-fit (GOF) test, Fourier ring correlation [28], and single-image SNR [29] to evaluate optimal choices for denoising. Quantitative analysis strikes the best balance between noise suppression and signal preservation. Here, noise suppression is based on how accurately the denoising algorithm for a particular set of parameters would obtain an estimate of the underlying noise distribution. The signal preservation is demonstrated by using "one-click" threshold based segmentation in the case of room temperature tomograms and using template-based search and segmentation of individual ribsomes in the case of cryogenic tomograms. The overall approach of using regions from tomograms that represent the background, and the strategy of iterating over different denoising algorithms is outlined in Figure 29.

For the KL-distance based GOF test, we consider $\boldsymbol{Y}$ as the noisy tomogram and $\boldsymbol{X}$ as the denoised tomogram. Consider that the tomogram contains n 2D slices. Hence, $\boldsymbol{Y} = \left(Y_i\right)_{i=1}^{n}$, where the index i represents the i^{th} 2D slice. Let the set of denoising algorithms containing m algorithms be denoted as $\boldsymbol{Z}$, $\boldsymbol{Z} = \left(Z_j\right)_{j=1}^{m}$, where the index j represents the j^{th} denoising algorithm. Now, $\boldsymbol{N = Y - X}$ represents the tomogram noise assuming that the noise is additive [110]. Our aim is to compare the noise present in the $\boldsymbol{Y}$ subset denoted as $\hat{\boldsymbol{Y}}$, shown in Fig. 1, where no biological data exists, to the same subset in $\boldsymbol{N}$, denoted as $\hat{\boldsymbol{N}}$ which we will model with various denoising algorithms. For example, in the *Bdellovibrio* tomograms, $\hat{\boldsymbol{Y}}$ would be the region corresponding to vitreous ice outside of the cell. We then evaluate different denoising algorithms by comparing each 2D slice of $\hat{\boldsymbol{Y}} = \left(\hat{Y}_i\right)_{i=1}^{n}$ compared statistically with $\hat{\boldsymbol{N}} = \left(\hat{N}_i\right)_{i=1}^{n}$. We use a non-parametric goodness-of-fit test based on the Kullback-Leibler distance to evaluate

statistically how close in distribution the noise present in $\hat{\boldsymbol{Y}}$ compares with $\hat{\boldsymbol{N}}$. Hence, our aim to is evaluate

$$\min_{\boldsymbol{Z}} \left\{ \hat{Y}_i - \hat{N}_i \right\}_d, \tag{65}$$

where the minimization is over the set of denoising algorithms $\boldsymbol{Z} = \left(Z_j \right)_{j=1}^{m}$ and over the valid parameter space for each algorithm. $\left\{ \hat{Y}_i - \hat{N}_i \right\}_d$ denotes the minimization in distribution. This minimization is evaluated using KL-distance based goodness-of-fit test, as detailed below.

Let $S_{\hat{Y}_i}$ and $S_{\hat{N}_i}$ be the vector of noise samples from $\hat{Y}_i$ and $\hat{N}_i$ respectively obtained from each 2D slice. We form the following null hypothesis:

$$H_0 : S_{\hat{Y}_i} \text{ and } S_{\hat{N}_i} \text{ follow identical distributions.} \tag{66}$$

The alternate hypothesis is that, statistically, the two noise samples do not follow the same distribution. Even though the underlying distributions are non-Gaussian, a non-parametric GOF test can reject the null hypothesis when the samples are statistically different. Here, the *p-value (p),* which lies between 0 and 1, is the lowest significance level denoted as β at which the null hypothesis can be rejected. The null hypothesis is often rejected if $p < 0.1$. In our case, we mostly deal with discrete distributions with large sample sets and hence conventional tests such as Kolmogorov-Smirnov (KS) and Kruskal-Wallis analysis-of-variance tests, which assume that the distributions are continuous and are quite sensitive to this assumption, cannot be applied here. Furthermore, the chi-square GOF test is based on a parametric assumption and is useful to compare known distributions, whereas we do not assume any a priori noise distribution and as a result, chi-square test is not a good test for non-parametric cases. To deal with this problem we constructed a robust non-parametric GOF test based on the Kullback-Leibler distance, or relative entropy [27]. The KL-distance based GOF test weights the difference of the two distributions across all bins by the probability mass at each bin rather than relying on maximum vertical variation at a single point, and hence is more suited for our discrete case.

For any two discrete probability mass functions (*pmf's*) t_1 and t_2, defined over the same set of values T, the KL-distance of t_1 relative to t_2 is

$$D(t_1 \,||\, t_2) = \sum_{k \in T} t_1(k) \log_2 \frac{t_1(k)}{t_2(k)} \,, \tag{67}$$

and $D(t_1 \,||\, t_2) \geq 0$, where equality holds if the two pmf's are identical. Note that this distance is asymmetric, i.e., $D(t_1 \,||\, t_2) \neq D(t_2 \,||\, t_1)$. The KL-test is performed as follows,

1. Estimate the *pmf's* $s_{\hat{Y}_i}$ and $s_{\hat{N}_i}$ where the bins are chosen according to the Freedman-Diaconis rule (see Appendix) from the samples $S_{\hat{Y}_i}$ and $S_{\hat{N}_i}$ respectively.
2. Calculate the KL-distance $D(s_{\hat{Y}_i} \,||\, s_{\hat{N}_i})$ between the true noise[1] relative to the estimated noise.
3. Let $\tilde{s}_{\hat{Y}_i}$ and $\overline{s}_{\hat{Y}_i}$ be the corresponding *pmf's* of any two randomly chosen partitions $\tilde{S}_{\hat{Y}_i}$ and $\overline{S}_{\hat{Y}_i}$ of the noisy samples $S_{\hat{Y}_i}$. Estimate the distribution of the KL-distance $D(\tilde{s}_{\hat{Y}_i} \,||\, \overline{s}_{\hat{Y}_i})$. If this process of partitioning is repeated a number of times, then the distribution of $D(\tilde{s}_{\hat{Y}_i} \,||\, \overline{s}_{\hat{Y}_i})$ can be estimated.
4. Reject the null hypothesis if $p \approx \mathrm{P}\left[D(s_{\hat{Y}_i} \,||\, s_{\hat{N}_i}) \leq D(\tilde{s}_{\hat{Y}_i} \,||\, \overline{s}_{\hat{Y}_i}) \right] < 0.1$ where $P[]$ is the probability operator.

As a second quantitative measure of denoising, we estimated single-image SNR using the technique outlined in [29], which assumes that the noise spectrum is white[2] and that the noise is uncorrelated from pixel to pixel. The SNR is given as [29]

$$SNR = \frac{\rho_{12}}{(1-\rho_{12})} = \frac{\phi_{11}^{NF}(0,0) - \mu_1^2}{\sigma_1^2 - (\phi_{11}^{NF}(0,0) - \mu_1^2)} = \frac{\phi_{11}^{NF}(0,0) - \mu_1^2}{\phi_{11}(0,0) - \phi_{11}^{NF}(0,0)} \,, \tag{68}$$

where $\phi_{11}(0,0)$ and $\phi_{11}^{NF}(0,0)$ are the autocorrelation functions (ACF) of the unprocessed and noise free (NF) images, respectively, and μ_1 and σ_1 are the mean and variance of the unprocessed image, respectively. Since we do not have a noise-free reference image, $\phi_{11}^{NF}(0,0)$ is estimated assuming that its value is the same as the ACF of the neighboring offsets, i.e.,

[1] The true noise samples are from the recorded tomograms at room or cryogenic temperatures.
[2] White noise is a random signal (or process) that has a flat power spectral density.

$\left\langle \phi_{11}^{NF}(0,0) \right\rangle \approx (\phi_{11}(1,0) + \phi_{11}(0,1))/2$. This estimate is reasonable if the ACF changes slowly at the origin.

A third and last measure of denoising quality was obtained using Fourier ring correlation (FRC) as a tool to compare the similarity of correlation coefficients in Fourier rings for two statistically independent sets using:

$$FRC(k,\Delta k) = \frac{\mathrm{Re}\left\{ \sum_{[k,\Delta k]} F_1(\boldsymbol{k}) F_2^*(\boldsymbol{k}) \right\}}{\left\{ \sum_{[k,\Delta k]} |F_1(\boldsymbol{k})|^2 \sum_{[k,\Delta k]} |F_2(\boldsymbol{k})|^2 \right\}^{1/2}}, \tag{69}$$

where $F_1(\boldsymbol{k})$ and $F_2(\boldsymbol{k})$ are the discrete Fourier transforms of the two subset averages, with the spatial frequency $\boldsymbol{k}$. Δk is the ring width and it is assumed that all values on the regular Fourier grid (k_x, k_y) are within the Nyquist range [28].

4.3 Feature Identification by Automated Segmentation

Previously reported methods for segmentation have generally fallen into two categories, namely region-based and contour-based approaches (see [111] and references therein). As pointed out in [112], local pixel operations are sensitive to the noise levels found in tomograms, and contour-based methods are prone to distortions since these methods use information from the local gradient. The work presented by Frangakis and Hegerl [112] deals with a segmentation procedure based on optimization of global cost function which uses pixel properties such as gray level and proximity in conjunction with an eigenvector analysis and hence does not require any user interaction. Sandberg et al.[113] have combined Line Filter Transform (LFT) and the Orientation Filter Transform (OFT) with contour extraction and labeling method to segment thin structures in 2D images in each slice of the tomogram. A shape-based segmentation technique is applied for segmenting microtubules by Jiang et al [114]. Since the fidelity of the segmentation in these different examples relies on the denoising procedures applied prior to segmentation, we used two different approaches to evaluate improvements in image interpretability by denoising.

For the HIV-infected macrophages, we compared surface spikes identified manually in the tomograms with those identified by automated one-click segmentation (i.e., using a single threshold value) on the same tomograms (after denoising) in the environment of the visualization program Amira [115], which executes a simple procedure of 3D density threshold region-growing from user-marked voxels and in a user-defined volume range. In addition to determining that spikes were correctly identified, we compared the estimated volumes of surface spikes obtained by automated segmentation with those obtained by careful user-guided manual segmentation.

For automated detection of ribosome-like particles in *Bdellovibrio* tomograms, we used a template-based strategy. We note that our automated segmentation procedure does not assume any knowledge of the biological features that are segmented or the nature of the search space. To reduce detection bias, we used spherical templates of varying radii, coupled with a constraint on the range of acceptable density values to search the volume. We located particles by finding the values that minimize the following error integral,

$$\min_{\alpha,t}\int_{C}\left(\alpha(t)g(x)-f(x-t)\right)^2 dx=\min_{\alpha,t}\int\left(\alpha(t)g(x)-f(x-t)\right)^2 h(x)\,dx, \qquad (70)$$

where $g(x)$ is the template (uniform sphere in our case), $f(x)$ is the given reconstructed tomogram, region C is a sphere slightly bigger than the template where we compare the template with the sub-volume, $h(x)$ is a mask for that region; $h(x)=\begin{cases}1 & x\in C\\ 0 & x\notin C\end{cases}$, t is the translation parameter and α is a coefficient that might vary for each t and allows the data to have a scaled version of the template. We also specify a threshold $\hat{\alpha}$ which is the minimum acceptable scaling coefficient. The above function is convex with respect to α and hence there is only one minimum which can be found by equating the derivative to zero as,

$$2\int g(x)\left(\alpha(t)g(x)-f(x-t)\right)h(x)\,dx=0. \qquad (71)$$

After the change of variable $y=x-t$ and using the fact that $g(x)$ is an even function (we consider a radially symmetric template), we have

$$\alpha(t)=\frac{\int g(x)f(x-t)h(x)\,dx}{\int g^2(x)h(x)\,dx}=\frac{\int g(-y-t)f(y)\,dy}{\int g^2(x)\,dx}=\frac{1}{n}v(t), \qquad (72)$$

where $n = \int g^2(x)dx$ and $v(t) = (f * g)(-t)$. Note that we can always write the product $g(x)h(x)$ as $g(x)$, because the non-zero part of *g* is included in the non-zero part of the mask. Therefore, the problem is reduced to

$$\min_t \int \left(\frac{1}{n} v(t) g(x) - f(x-t) \right)^2 h(x)dx . \quad (73)$$

The minimization leads to

$$\min_t D(t) = \min_t \left(c(t) - \frac{1}{n} v^2(t) \right); \alpha(t) = \frac{1}{n} v(t) \geq \hat{\alpha} , \quad (74)$$

where $c(t) = \left(f^2 * h \right)(-t)$. $D(t)$ is the least squared error between the scaled template and the reconstructed tomogram. We can either select all the eligible voxels where $D(t)$ is greater than a predetermined threshold value, or sort all the eligible voxels, based on $D(t)$, and select the *M* top voxels as the centers of the ribosome-like particles and manually find the best *M*. Once a point is chosen as the center of a ribosome, all voxels around it that are closer than the assumed diameter of the particle are marked, and not considered again in the search. We note that the feature extraction method we use for ribosome detection only has a computational complexity of *O(NlogN)*, where *N* is the number of voxels, and is therefore effective for the purposes of automated segmentation.

4.4 Results and Discussion

The goal of denoising is to remove as much noise as possible by varying the parameters over a wide range while ensuring that biologically relevant features are retained (Figure 29). Starting with a stack of 2D images that constitute the initial unprocessed tomogram, we used a variety of denoising strategies with the aim of obtaining the best possible separation of signal from noise, and then evaluated the performance of each algorithm using the quantitative tests described in previous section. Two types of test tomograms were used in our study. One set of tomograms that was used was obtained from fixed, plastic-embedded HIV-1 infected macrophage specimens imaged at room temperature and at 120 kV. Dual-axis tilt series were obtained and reconstructed

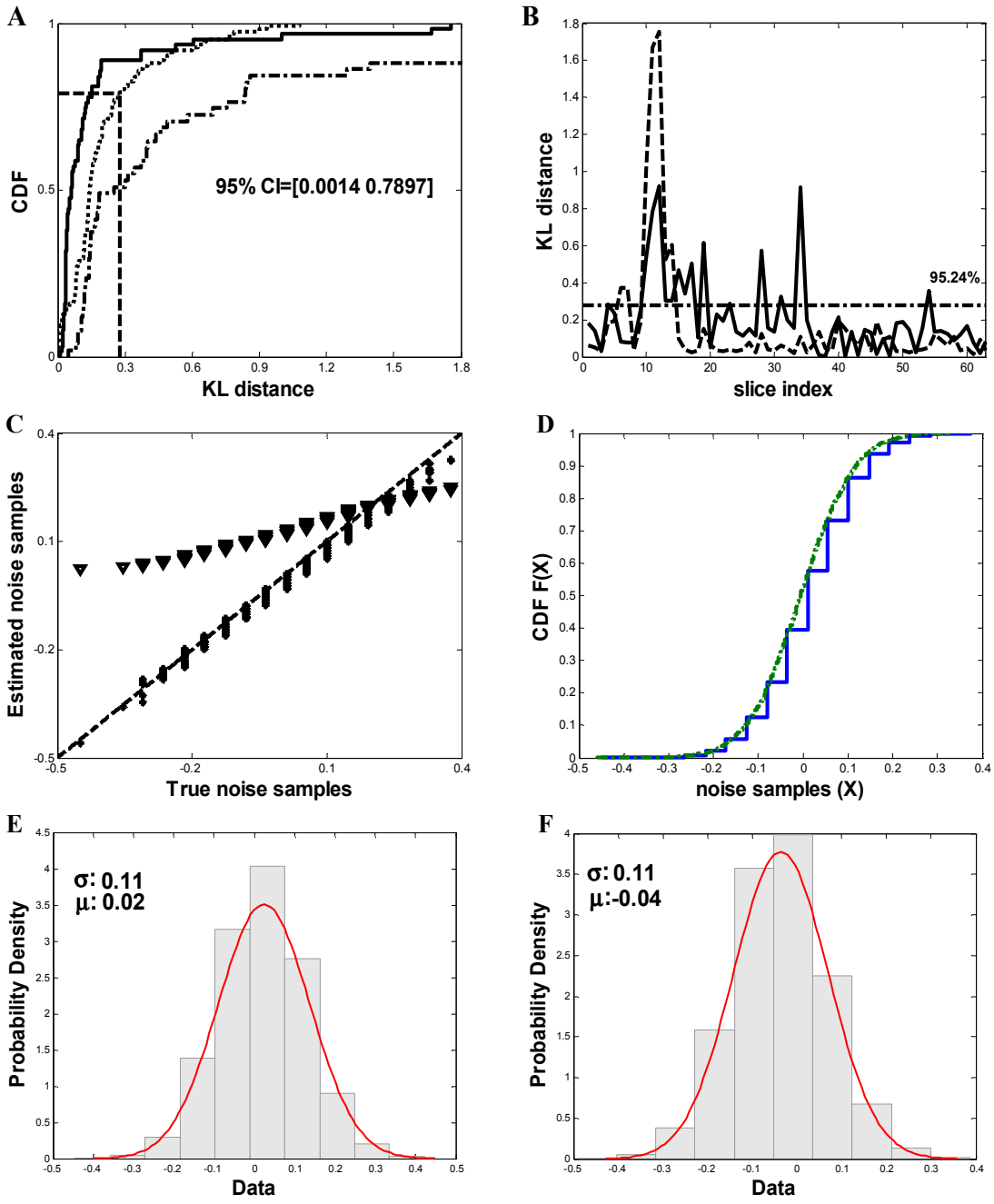


Figure 30. Denoising analysis of frozen-hydrated *Bdellovibrio* cell tomogram from data recorded at liquid nitrogen temperatures (A) Comparison of the cumulative distributions of the KL-distance $D(\tilde{s}_{\hat{Y}_i} \,||\, \overline{s}_{\hat{Y}_i})$ shown as dotted line (·····)($\tilde{s}_{\hat{Y}_i}$ and $\overline{s}_{\hat{Y}_i}$ are the corresponding probability mass functions of any two randomly chosen partitions $\tilde{S}_{\hat{Y}_i}$ and $\overline{S}_{\hat{Y}_i}$ of the noisy samples $S_{\hat{Y}_i}$) and $D(s_{\hat{Y}_i} \,||\, s_{\hat{N}_i})$ shown as solid line (——) respectively ($s_{\hat{Y}_i}$ and $s_{\hat{N}_i}$ are the probability mass functions of $S_{\hat{Y}_i}$ and $S_{\hat{N}_i}$) from the tomogram after denoising with NAD. The 95% confidence intervals are shown. The dashed-dotted line represents the cumulative distribution of the KL-distance using Wiener filtering. (B)

Comparison of KL-distances. The solid curve (——) represents the KL-distance between random partitions of the raw noise samples and the dotted line (·····) represents the KL-distance between true and the estimated noise samples. The horizontal dashed-dotted line depicts the upper confidence values. (C) Quantile-quantile (q-q) plot of true and the estimated noise samples after denoising using NAD denoising algorithm with optimal parameters (∗) and using Wiener filtering (∇). The 45° slope line is shown in the plot as a dashed line (----). (D) Comparison of the distributions of true (——) and the estimated noise samples (·····) after denoising. (E) & (F) Comparison of the probability density functions of true and estimated noise samples. A Gaussian fit is computed with the mean and variance shown in the respective plots.

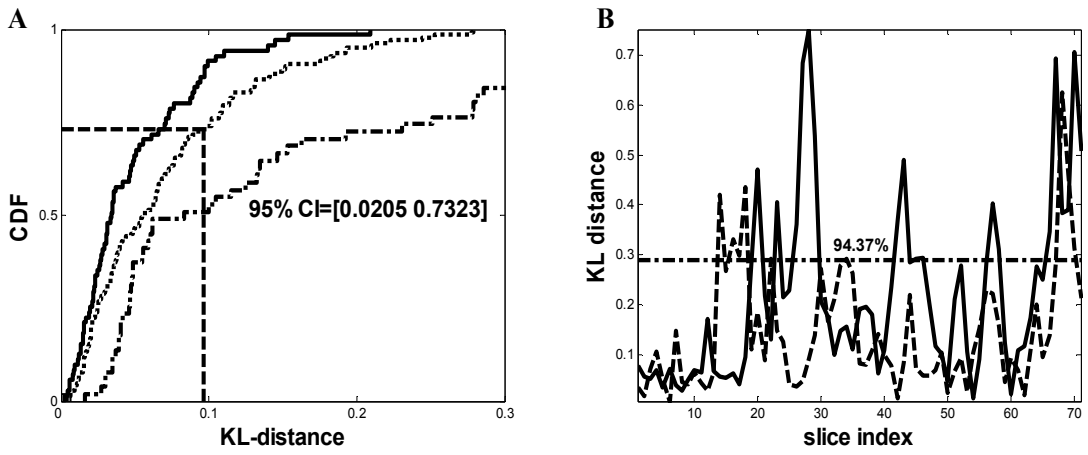


Figure 31. Denoising analysis of tomogram of plastic-embedded HIV-1 infected macrophages obtained from data recorded at room temperature. (A) Comparison of the cumulative distribution of the KL-distance $D(\tilde{s}_{\hat{Y}_i} \,||\, \overline{s}_{\hat{Y}_i})$ shown as dotted line (·····) and $D(s_{\hat{Y}_i} \,||\, s_{\hat{N}_i})$ shown as solid line (——) respectively for a dual-axis tomogram using the phase-preserving denoising algorithm. The 95% confidence interval is shown in the figure. The dashed-dotted line represents the cumulative distribution of the KL-distance using median filtering. (B) Comparison of KL-distances. The solid curve (——) represents the KL-distance between random partitions of noisy samples and the dotted line (·····) represents the KL-distance between true and the estimated samples. The horizontal line depicts the upper confidence values.

using SIRT-based reconstruction algorithms [116] to obtain tomograms. SIRT generally approaches convergence after a high number of iterations (often more than 500); however, the technique as implemented in FEI's Inspect3D software results in a contrast maximum at around 17 iterations. This also reduces required computation time.

For the purpose of high-resolution structure determination in cryo-electron tomography, it is critical that convergence is approximated. However, with heavy-metal-stained, plastic-embedded sections, the resolution is irrelevant below about 10 nm and the purpose of tomography is gross morphological characterization. Therefore, 17 iterations is an arbitrary yet valid number of SIRT iterations for comparison of denoising techniques using tomograms of plastic-embedded sections. A second set of tomograms used was obtained from plunge-frozen *Bdellovibrio* cells

imaged at liquid nitrogen temperatures at 300 kV and reconstructed using weighted back-projection algorithms as implemented in IMOD [56]. One should note that we compute the KL divergence slice by slice and only pick the "optimal" algorithm when out of the total number of slices in the tomogram at least 95% is above the confidence interval. We do it on a per slice basis for the following reasons:

(a) This provides a robust way of comparing two discrete distributions

(b) The computational overhead is significantly reduced.

Also, due to the asymmetric property of KL-distance, we average them.

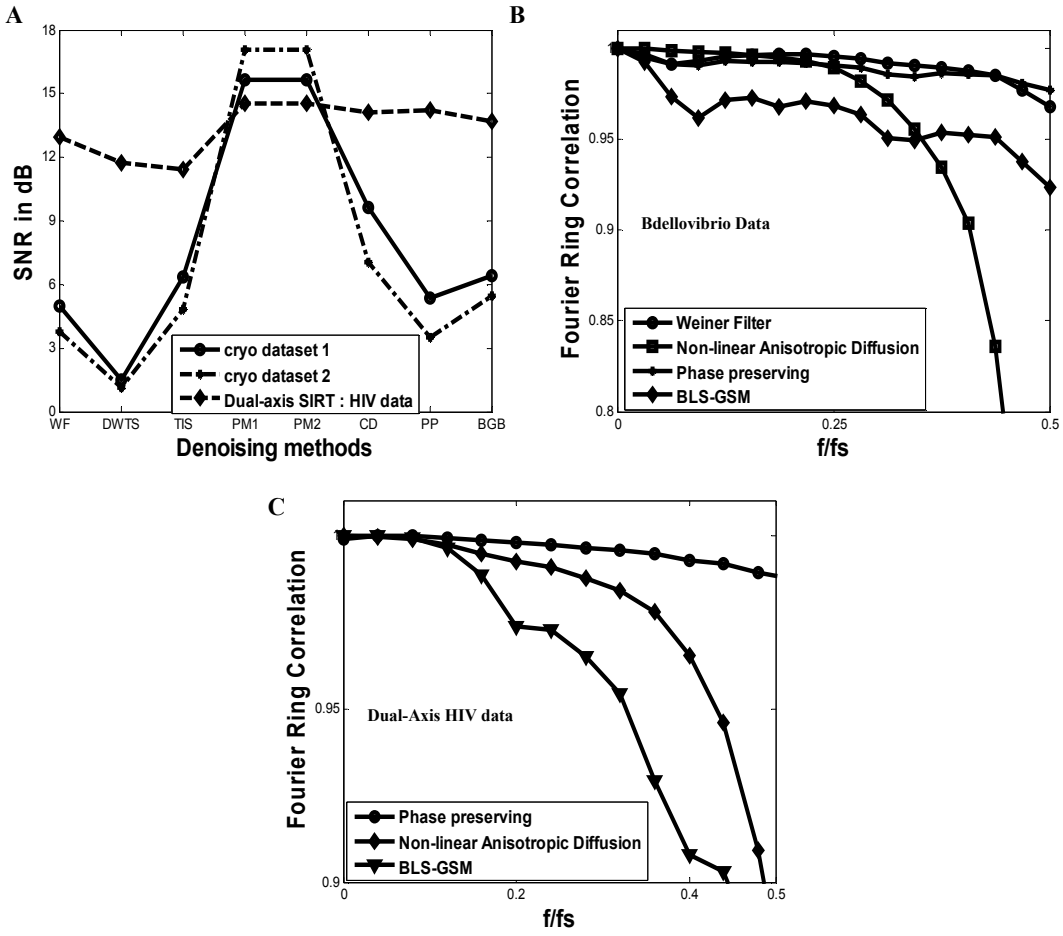


Figure 32. (A) Single-image (signal-to-noise ratio) SNR testing method using various denoising methods for the *Bdellovibrio* (shown as —●— and -·-·*-·-·) and *HIV-1* (shown as ---♦---) tomograms recorded at cryogenic and room temperatures, respectively. (B & C) Fourier Ring Correlation (FRC) plots. Comparison of the performance of different denoising algorithms on *Bdellovibrio* (B) and *HIV-1* (C) tomograms. Here, $\frac{f}{f_s}$ refers to the normalized spatial frequency.

In Figure 30, Figure 31, and Figure 32, we present the performance of different denoising algorithms on these two data sets as assessed by three tests: the KL-distance test (Figure 30 and Figure 31), single image SNR analysis (Figure 32), and the frequency dependence of Fourier Ring Correlation (Figure 32). These findings are described in more detail below.

In Figure 30(A), we present a comparison of the cumulative distribution function ($F(\Theta) = P(\Theta \leq \theta)$, which gives the probability that a random variable Θ is less than θ) of the KL-distance between the true and the estimated noise samples ($D(s_{\hat{Y}_i} \| s_{\hat{N}_i})$) that are compared against the KL-distance of two random partitions of true noise samples ($D(\tilde{s}_{\hat{Y}_i} \| \overline{s}_{\hat{Y}_i})$) obtained from tomograms of plunge-frozen *Bdellovibrio* cells denoised using NAD and Wiener filtering. As shown in the plot, the *p-value* of the KL-distance based goodness-of-fit test, using the 95% ($1-\beta = 0.95$) confidence interval is estimated at 0.9524 for denoising using NAD and 0.5071 for denoising using Wiener filtering. The performance of diffusion equation of Weickert [107] yielded a *p-value* of 0.6923. We also found that the performance of the bilateral filter in terms of denoising on room and cryogenic temperature data was below phase preserving and NAD algorithms respectively. The plot of the KL-distance of $D(s_{\hat{Y}_i} \| s_{\hat{N}_i})$ and $D(\tilde{s}_{\hat{Y}_i} \| \overline{s}_{\hat{Y}_i})$ as a function of the slice index is shown in Figure 30(B). From the figure, we can observe that 95.24% of the data points are < 0.277, which is the upper confidence value. These results indicate that the NAD-based denoising does an excellent job of estimating the noise distributions in the data, in contrast to denoising using a Wiener filtering algorithm. For further analysis of noise estimation using NAD, we compute the quantile-quantile *(q-q)* plot (Figure 30(C)), which is a graphical technique for determining if two data sets come from populations with a common distribution. If the two sets come from a population with the same distribution, the points should fall approximately along a 45° reference line. The q-q plot between the true and the estimated noise samples for *Bdellovibrio* data denoised using NAD algorithm falls close to this 45° line, while that obtained using the Wiener filtering algorithm deviates significantly. The plot of the cumulative distribution function of the estimated and the true noise samples is shown in Figure 30(D), where it can be observed that the two curves overlap. Yet another quantitative measure of the performance of the NAD algorithm is shown in comparison of the probability density

functions of the true and estimated noise sets (Figure 30(E), & Figure 30(F)) demonstrates that their respective variances are similar to each other. Together, these results provide confidence that the NAD method that we have employed can provide a reliable approach to estimating noise in the tomograms obtained from vitrified *Bdellovibrio* cells.

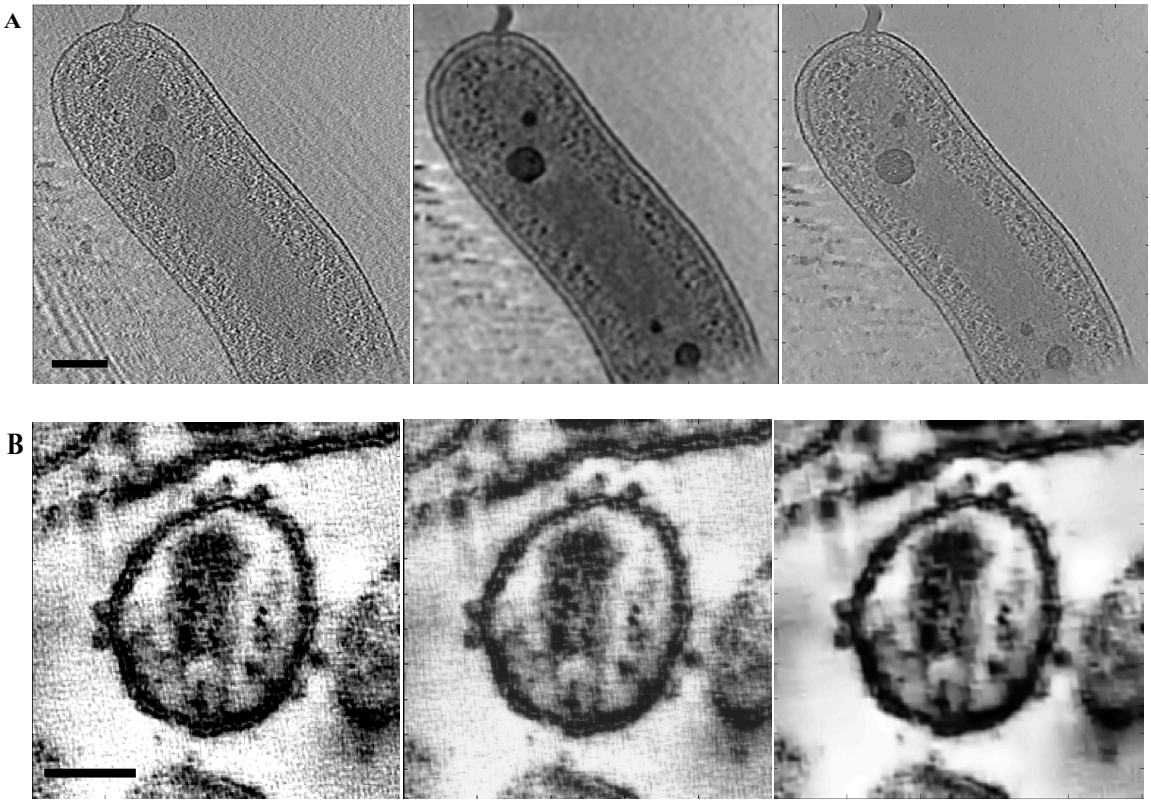


Figure 33. Denoising of single 2D slice of *Bdellovibrio* (A) and *HIV-1* (B) tomograms using nonlinear anisotropic diffusion (NAD) and phase preserving (PP) algorithms respectively. In each case, we show the unprocessed image, followed by the same image denoised with NAD, and PP algorithms, respectively. The NAD algorithm provides better detection of ribosomes in the case of the Bdellovibrio tomogram, while the PP algorithm provides better detail of surface spikes in the case of the HIV-1 tomogram. Scale bar in panels A and B are 100nm and 50nm, respectively.

The results of a similar analysis using tomograms recorded from HIV-1 infected macrophages obtained at room temperature and the results of the comparison of $D(s_{\hat{Y}_i} \| s_{\hat{N}_i})$ and $D(\tilde{s}_{\hat{Y}_i} \| \overline{s}_{\hat{Y}_i})$ are presented in Figure 31(A), which illustrates that in contrast to the findings with the cryo tomographic data, the phase preserving algorithm estimated the noise distributions in the data better compared to the rest of the algorithms. A comparison against median filtering is shown using a dashed-dotted line in Figure 31(A). The respective plot of the KL-distance of

$D(s_{\hat{Y}_i} \| s_{\hat{N}_i})$ and $D(\tilde{s}_{\hat{Y}_i} \| \overline{s}_{\hat{Y}_i})$ as a function of the slice index is shown in Figure 31(B) where 94.37% of the data points are less than 0.2902, which is the upper confidence value. The results in Figure 30 and Figure 31 show that using the NAD algorithm on the *Bdellovibrio* tomogram, and the phase preserving algorithm on the room temperature data sets, resulted in estimated noise samples ($\hat{N}$) that are indeed close in distribution to those of true noise samples.

As an independent measure of the performance of the different denoising methods, we compared the SNR of each denoised 2D slice for the tomograms obtained at cryogenic and room temperatures (Figure 32(A)). The figure depicts an average SNR computed over the whole tomogram. As in the case of KL divergence, this gives an estimate of the performance of a denoising algorithm on a per slice basis. The plots confirm that the NAD methods produce the highest SNR for the cryo tomograhic data, confirming the results from the KL-distance plots. For the room temperature data from HIV-1 infected macrophages, the SNR analysis suggests that performance of both the NAD and phase preserving approaches is comparable.

We also computed the Fourier ring correlation (FRC) [28] to compare various denoising methods (Figure 32(B), Figure 32(C)), implemented based on Equation (69). The complex Fourier coefficients are calculated from the unprocessed and the processed ensemble. Here, improvement in resolution corresponds to an improvement in SNR. It can be observed that in the case of *Bdellovibrio* tomograms (Figure 32(B)), the NAD algorithm exhibits good performance at low frequencies, and due to the low-pass characteristic of the diffusion equation, higher frequency components ($> 0.5 f_s$) of the signal are degraded. However, the important point is that the performance of the NAD algorithm is slightly better than the performance of phase preserving and the Wiener filter algorithms at the lower frequencies that matter the most for interpreting lower resolution architectural information from electron tomograms of biological specimens. Similarly, in the case of the room temperature tomograms, the FRC plots (Figure 32(C)) show that the performance of the phase preserving algorithm is superior to NAD at most frequencies, thus confirming the findings from the KL-distance plot analysis in Figure 31. A principal conclusion from the above analyses demonstrate that the noise distribution and statistics vary for tomograms recorded at room and cryogenic temperatures and hence the choice of denoising algorithm needs to be guided by the type of noise present in the individual data sets. Illustration

of the denoising with optimal, as well as sub-optimal denoising of representative *Bdellovibrio* and HIV-1 tomograms is shown in Figure 33 we compare NAD and phase preserving algorithms. Comparison of the unprocessed and denoised slice of a second *Bdellovibrio* tomogram using NAD is shown in Figure 34.

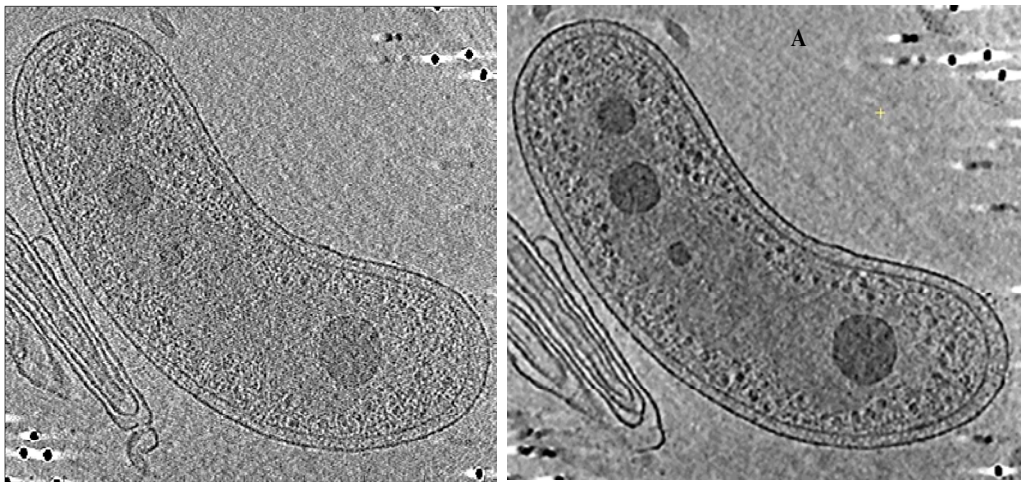


Figure 34. Denoising of single 2D slice of *Bdellovibrio* (A) tomograms using nonlinear anisotropic diffusion

A practical test of the value of denoising is the accuracy of automated segmentation relative to manual segmentation of the denoised tomograms carried out by an experienced user. In the *Bdellovibrio* tomograms, the distribution of particles in the cytoplasm that are most likely ribosomes is clearly evident in the tomograms after denoising. We therefore compared differences between manual identification of these putative ribosomes and the automated identification of these particles using unbiased spherical templates with different radii. The putative ribosomes are extracted based on intensity, which needs to be greater than a specified threshold, and least-squares minimization of the error between the scaled template and the reconstructed tomogram. The ribosome-like particles extracted from the tomogram by template matching are sorted based on the error between the scaled template and the features extracted from the tomogram to obtain the best matches. As the template, we used a uniform sphere of radius 4 pixels (corresponding to 22 nm; roughly the size of a ribosome); this radius minimized false positives and provided the best matches when compared against manual segmentation.

Using automated segmentation, we extracted 419 putative ribosomes from the *Bdellovibrio* tomogram shown in Figure 35, and 437 putative ribosomes derived by manual

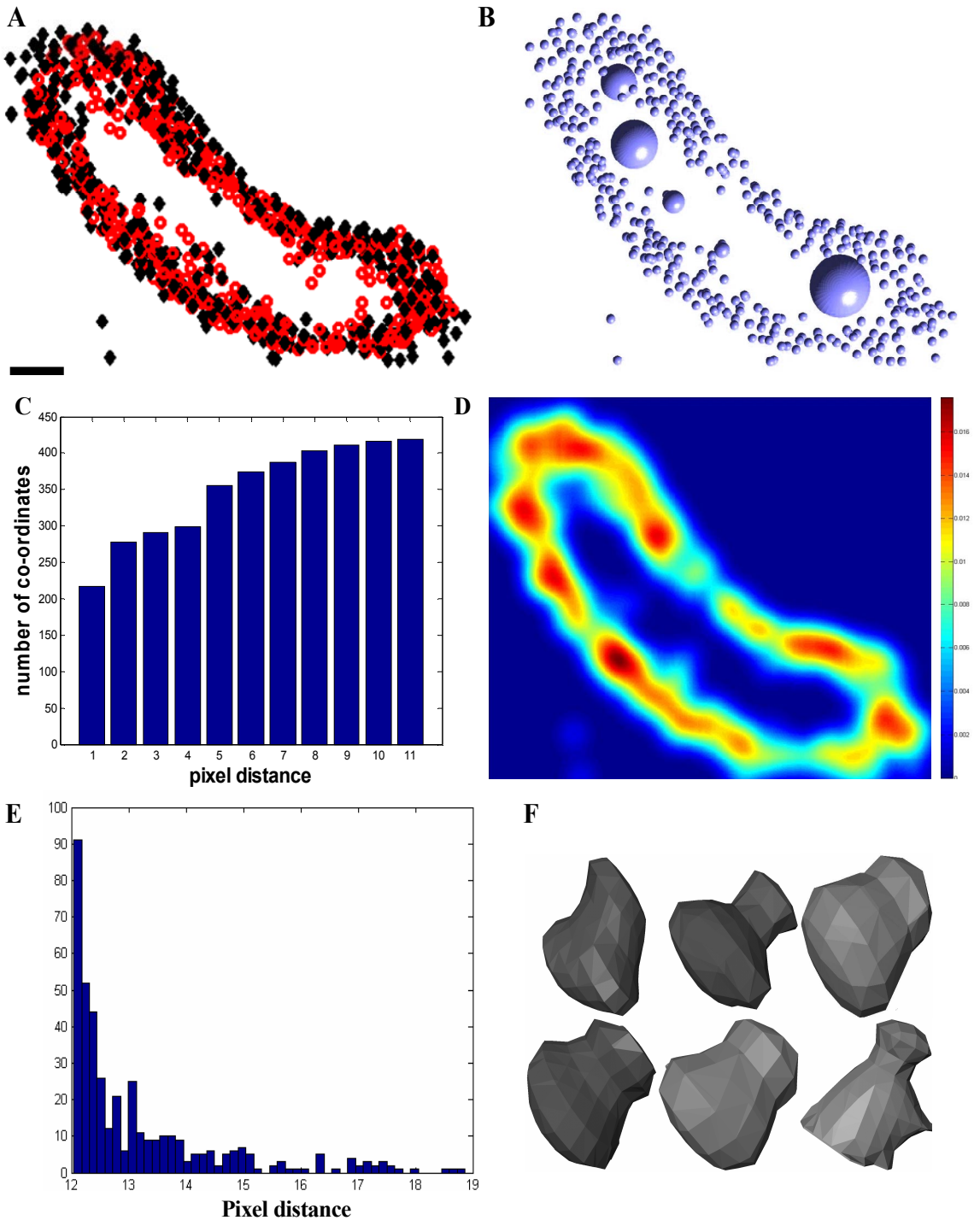


Figure 35. (A) Comparison of the locations of the extracted ribosome-like particles from *Bdellovibrio* data using manual (o) and automated (◊) segmentation (B) locations extracted using only automated segmentation. (C) Euclidian pixel distance between the semi-automated and automated segmentation. (D) Ribosome density where the different colors in the bar indicate the density values (Scale bar: 100nm). (E) Distance to the nearest neighbor histogram and (F) Shape of a few randomly picked ribosome-like particles extracted from the volumes of the denoised *Bdellovibrio* tomograms. Scale bars are 100 nm.

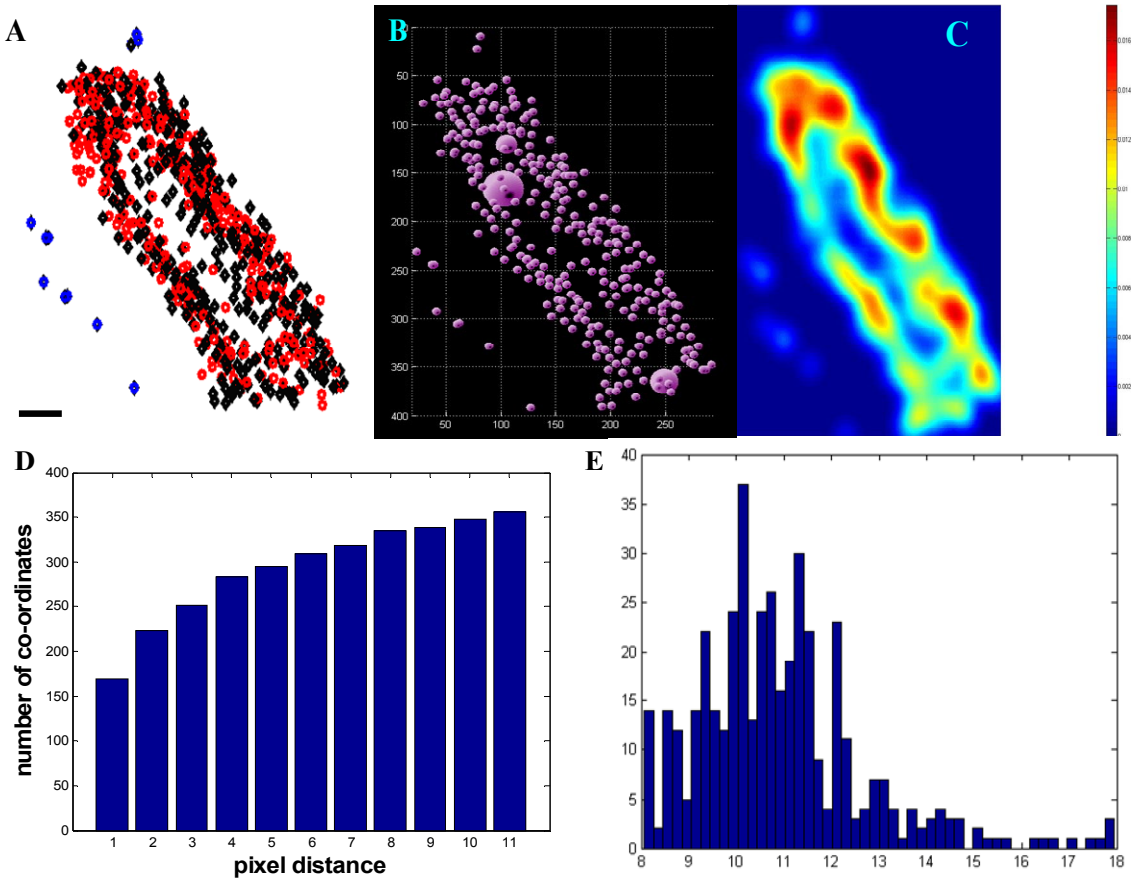


Figure 36. (A) Comparison of the locations of the extracted ribosome-like particles from *Bdellovibrio* data using manual (o) and automated (◊) segmentation (B) locations extracted using only automated segmentation. (C) Ribosome density where the different colors in the bar indicate the density values. (D) Euclidian pixel distance between the semi-automated and automated segmentation and (E) Distance to the nearest neighbor histogram. Scale bars are 100 nm

segmentation from the same tomogram. On the second dataset, we extracted 359 putative ribosomes shown in Figure 36, and 370 putative ribosomes derived by manual segmentation from the same tomogram. The centroids obtained using manual and automated techniques were compared and about 85% of the ribosomes were within 5 pixels, which is about half the ribosome size as shown in Figure 35(A) and Figure 36(A). About 92% of the ribosomes are within 7 pixels and the maximum error was within 11 pixels for all extracted volumes. The Euclidean pixel distance between the manual and automated techniques is shown in Figure 35(C) and Figure 36(D). Density of the ribosome-like particles and the histogram of the nearest neighbor distance are shown in Figure 35(D) & Figure 36(C) and Figure 35(E) & Figure 36(E) respectively. From

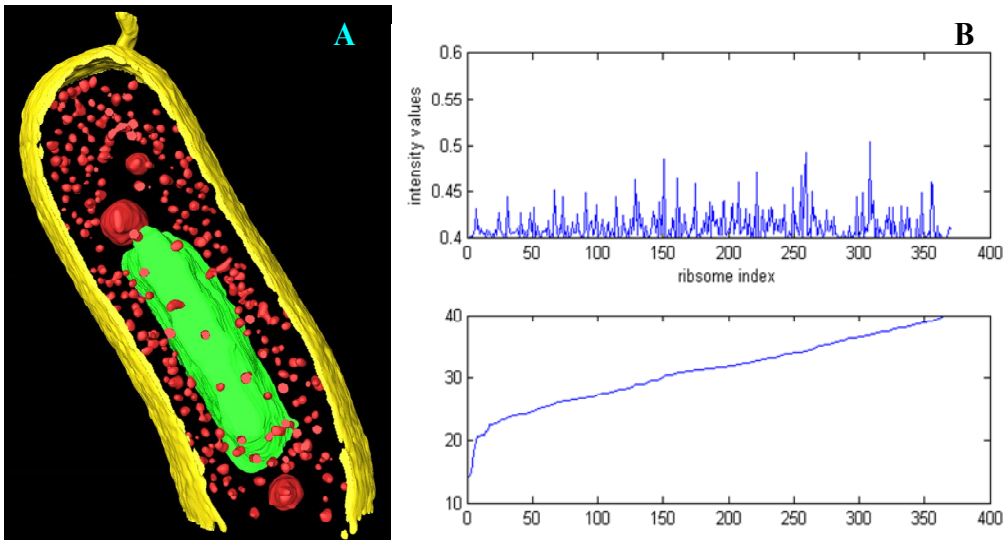
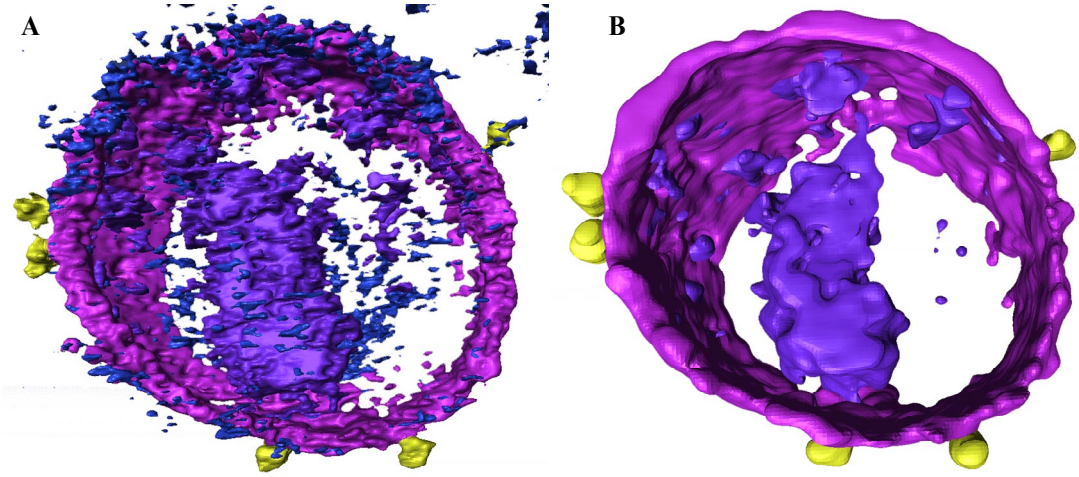


Figure 37. (A) Segmentation of denoised tomogram by an experienced user in the environment of Amira and (B) Intensity and error values for the extracted ribosomes using template matching.

Table 2. Quantitative analysis

	Dataset 1	**Dataset 2**
Nearest Neighbor distribution (pixels)	Mean - 10.3 Variance - 2.89	13.3376 5.773
Ribosome Distribution (Avg/nm²)	0.0024	0.0025
Average ribosome count	373.75	432
Automatic procedure	359	419



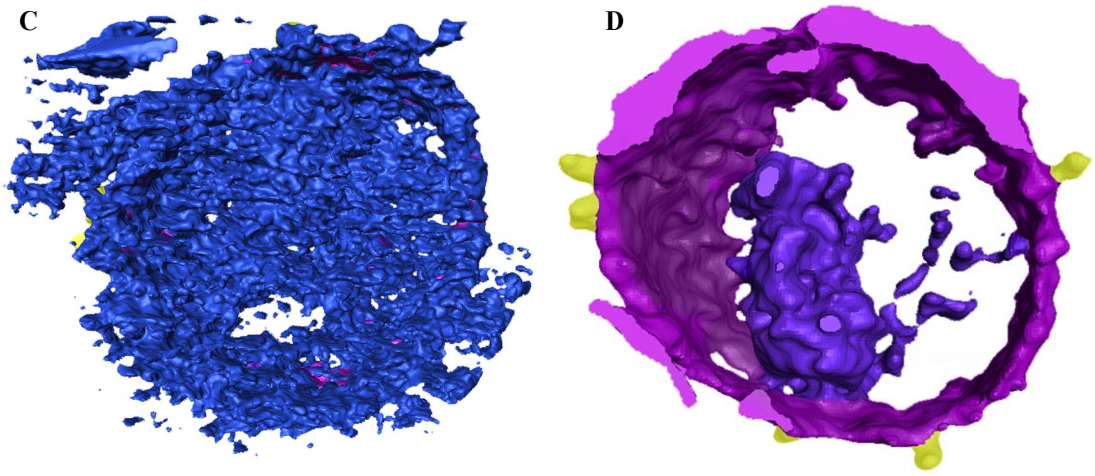


Figure 38. Comparison of feature extraction by manual vs. automated segmentation of dual-axis SIRT reconstructed HIV-1 tomogram. (A) unprocessed and (B) denoised tomogram visualized in the environment of Amira. (C) & (D) Comparison of "single-click" segmentation results for unprocessed and denoised tomograms respectively, with denoising performed using the phase-preserving algorithm

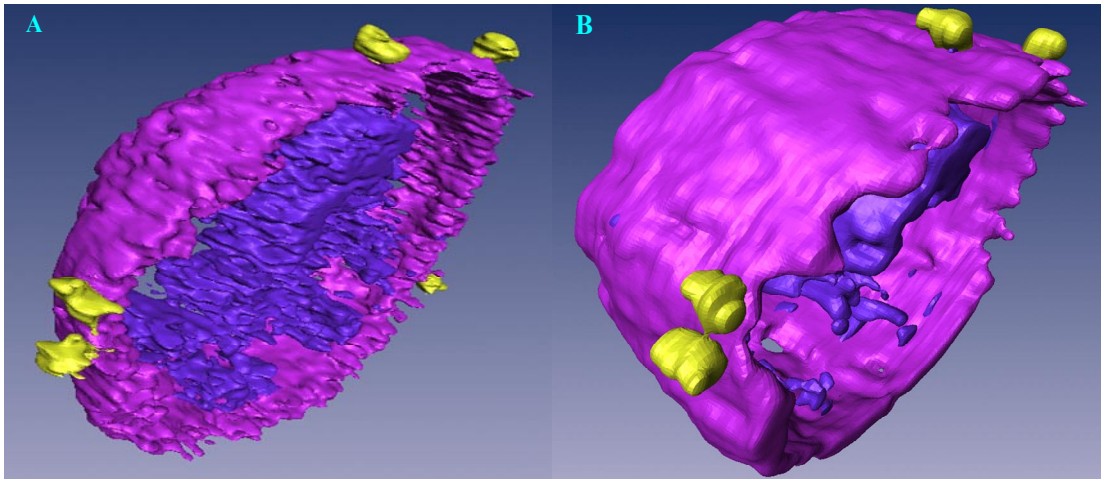


Figure 39. Comparison of denoising of dual-axis SIRT reconstructed HIV-1 tomogram: (A) Unprocessed and (B) denoised tomogram

the ribosome density (see Table 2) plots in Figure 35(D) and Figure 36(C), the ribosome count in the tomograms can be calculated as 432 and 374 compared to 419 and 359 that were obtained by automatic segmentation (see Table 2). Manual segmentation of ribosomes from a denoised tomogram by an experienced user is shown in Figure 37(A) and the plots of the matched intensities and error values between the template and the denoised tomogram is shown Figure 37(B). From Table 2, the mean and variance of the nearest neighbor distance are (13.3, 5.77) and (10.3, 2.89) pixels respectively. We also extracted the shapes of randomly picked putative

ribosomes from the denoised cryo-tomograms. Although noisy, the shapes of these randomly picked volumes are reminiscent of the expected bi-lobed shape of ribosomes.

For the room temperature tomograms of HIV-1 infected macrophages, we used the detection of the location and shapes of the surface spikes on the virus as a measure for comparing the performance of automated segmentation of denoised tomograms. The semi-automated segmentations, in which user-designated regions of the tomogram are subjected to varying thresholds in the case of the HIV-1 tomogram, are shown in Figure 38(A) and Figure 38(B) for raw and denoised tomograms, respectively, using the phase preserving algorithm (see Figure 39 for a side view). We then compared the results of "single-click" segmentation for dual-axis SIRT (DA-SIRT) reconstructed HIV-1 tomograms on raw and denoised tomograms (Figure 38(C) and Figure 38(D)). Single-click segmentation of a noisy DA-SIRT tomogram shown in Fig. 8C returned a surface that was highly non-contiguous and included noise. After the tomogram was denoised, one-click segmentation Figure 38(C) yielded a surface with a complete membrane, distinct core, and envelope glycoprotein spikes that closely matched the location of the spikes by manual segmentation as shown in Fig. 8B. When the raw and denoised spike volumes were compared, the denoised spike was elongated, distinct and the connection to the viral membrane was well preserved unlike the raw version which was difficult to interpret clearly. We compared the fully and semi-automated approaches on the HIV-1 tomogram in terms of the average volume estimates of viral spikes. The average volume using the semi-automated approach was 75 ± 7 $nm^3$and using fully automated method was 70 ± 18 nm^3 (see Table 3 for details). The results demonstrate that iterative denoising procedure coupled with quantitative and qualitative analysis facilitates automated segmentation of the volumes to efficiently extract information from large datasets in biological tomography.

Table 3. Volume occupied by the viral spikes

Spike index	Voxel size (Semi-Automated)	Voxel size (Automated)	Relative Error
1	942.6560	825.36	0.1244
2	757.5100	678.92	0.1037
3	962.9190	1209.40	0.2560
4	849.7470	691.77	0.1859

4.5 Summary

Electron tomograms are intrinsically noisy. This poses significant challenges for image interpretation, especially in the context of low dose and high-throughput data analysis. Our goal has been to evaluate the relative performance of different denoising methods in further improving the SNR, and to test whether these denoised tomograms can be processed automatically to extract biologically relevant information. We show here that denoising significantly improves the fidelity of automated feature extraction. The NAD algorithm performs best for recovering structural information from low-dose cryo tomograms, and spatial information such as ribosome distribution can be obtained automatically from denoised tomograms with results closely matching those obtained using semi-automated approaches. We show that molecular information such as the average volume of individual HIV viral spikes can be obtained from these denoised tomograms, and that the values closely match those obtained using manual user-guided approaches for room temperature tomograms of HIV-1 infected macrophages. The use of these valuable computational tools provides a further step for quantitative analysis of 3D structures determined using electron tomography.

CHAPTER 5

AUTOMATIC TEXTURE-BASED SEGMENTATION OF MITOCHONDRIA IN MNT-1 CELLS USING MACHINE LEARNING APPROACH

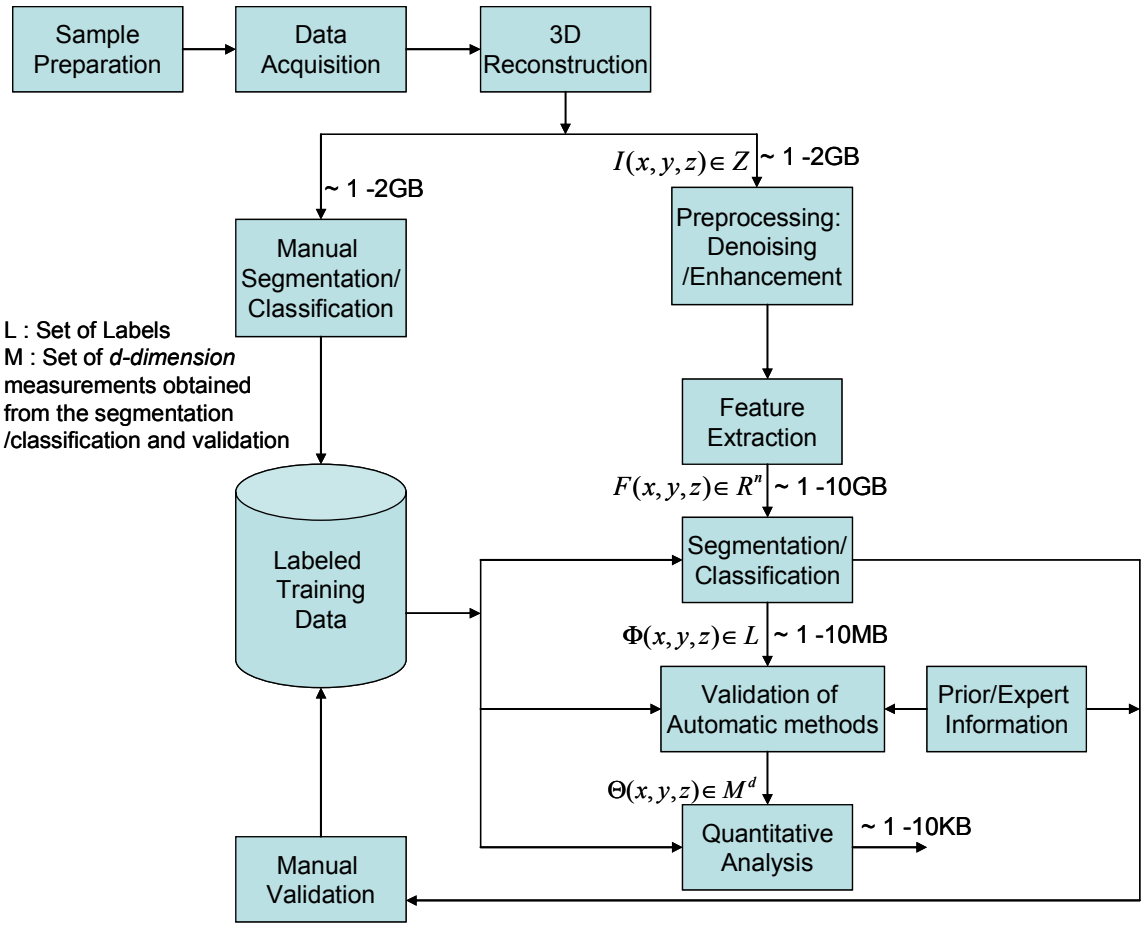


Figure 40. Block diagram of mining relevant byte-sized information from gigabyte sized tomograms using machine learning approach (modified figure from the NCI, NIH proposal).

This chapter presents an approach for automatic texture-based segmentation of mitochondria in MNT-1 cells imaged using an ion-abrasion scanning electron microscope. For cancer detection, a number of human melanoma whole cell tomograms (each 3D tomogram is about ~2GB) needs to analyzed. Hence, automatic tools that have minimal user intervention needs to be developed for high-throughput data mining and analysis. Challenges for such a tool in

electron tomography arise from low contrast and signal-to-noise ratio (SNR), appearance, geometry and viewpoint variation. The block diagram of the automatic procedure starting from sample preparation to quantitative analysis is shown in Figure 40. In this figure, we demonstrate the use of machine learning approach to mine relevant byte-sized information from gigabyte-sized tomograms. Section 2.6 of Chapter 2, provides background information about the importance of quantitative analysis of mitochondrial and Liposomal Doxorubicin formulations (Doxil) in cancer research. In particular, we investigate quantitative analysis such as volume of cytoplasm occupied by mitochondria, difference between surface area of inner and outer membranes and mean mitochondrial width which are indicative quantities of segregating a cancerous cell from a normal one in Section 5.5. Next, we test the robustness of our learning based segmentation approach on tomograms of Liposomal Doxorubicin formulations (Doxil), an anticancer nano-drug, imaged at cryogenic temperatures and study the radii and volume distribution which is of great interest in caner treatment.

Object recognition can be defined as the localization of object instances within an image. As mentioned in [117], object localization is to obtain class models that are invariant enough to incorporate naturally occurring intra-class variations and yet discriminative enough to distinguish between different classes. Using manually labeled ground truth images that represent object classes, the goal is to learn models such that novel instances of those objects in a new test image can be localized. In this work we focus on region-based segmentation (mitochondria in MNT-1 cell) using pixel wise classification of images into a trained list of regions. Both object and region localization usually are realized in practice using texton-based models that are developed in the context of texture recognition [118-120]. “Texton” and “visual word” have been used interchangeably in computer vision literature, which refers to clusters of filter responses in a high-dimensional space.

Particularly, textures are modeled as joint distribution of filter responses, where the distribution is represented as a frequency histogram of filter response cluster centers (textons) and the texture models are learnt from training images [117, 119, 121, 122]. Many texture-based and object category classification approaches have been applied to image databases, for e.g. CUReT [123], UIUCTex [124], CalTech101 [125], CalTech6 [126] and PASCAL challenge [127], scene

understanding [128], semantic image retrieval, web search and interactive image editing [117]. The classification algorithm is affected by parameters such as choice of filter bank and rotational invariance [117, 129-131], the size of texton dictionary as well as the number of training images (see [119] for a survey). Also in [132, 133], the authors have argued that classification can be obtained by using local pixel neighborhoods directly, without resorting to large scale filter banks. Image segmentation is also investigated based on Markov random field [41, 134], and incorporating classifiers in a conditional random field [135]. Texture recognition have been investigated under various factors such as lighting, viewing conditions [119, 129] brightness and color [136]. In [137], the authors separate texture from untextured regions, where contours are treated in the intervening contour framework and textures are analyzed using textons. They propose a graph theoretic framework of normalized cuts to find partitions of the image. Texture and object category classifications are studied using different key point detectors and descriptors, as well as different kernel and classifiers by Schmid et al [122]. Classification using support-vector machines (SVM) have also been extensively applied to object categorization [122], texture [138] and medical applications [139, 140]. A segmentation approach for retinal vessels based on textons has been investigated in [141]. Recently, multi-class boosting techniques have been applied to region-based classification of aerial imagery [128].

In biological tomograms, texture-based segmentation is challenging due to large size of the data, low contrast and signal-to-noise ratio (SNR), appearance, geometry and viewpoint variation. To the best of our knowledge, our work is one of the first to investigate automatic texture-based segmentation of mitochondria in MNT-1 cells imaged using ion-abrasion scanning electron microscope at resolutions better than ~6nm and ~20nm in the parallel and perpendicular, respectively, to the direction of ion beam milling. The block diagram of mining relevant byte-sized information from gigabyte-sized tomograms using learning approach is shown in Figure 40. The denoising algorithm implemented as a preprocessing step is based on variational filtering using an adaptive fidelity term that applies different levels of denoising in different regions [142]. This denoising method preserves the textures and smooth regions are better denoised in terms of signal-to-noise ratio compared to nonlinear-diffusion [96, 107, 110] and total variation methods

[143] that preserve sharpness and location of the edges, and in some cases enhance them, but significant small details and textured regions of the image are ignored.

Unlike other published work in ET [51], our approach is novel in the sense that it is based on automation of the entire process from data collection, reconstruction, denoising, feature extraction and averaging using statistical, information-theoretic, and learning tools. For cancer detection, a number of human melanoma whole cell tomograms (each 3D tomogram is about ~2GB) needs to analyzed. Various semi-automatic segmentation methods have been developed in the past for biological applications (see chapter, 30 of [9]) that require parameter setting or some sort of a priori information and hence would a major bottleneck for high-throughput data mining and analysis. Our goal here is to develop a machine learning tool that has minimal user intervention for high-throughput imaging. In our approach, user intervention is limited to choosing the parameters just for the training set (including the parameters for texture preserving filter), which is then retained for subsequent test datasets.

We have borrowed texture-based models developed in the context of texture [119] and object [117] classification. Given manually labeled images, that contain both segmentation maps and rough annotations of regions, our supervised learning algorithm learns the model such that it can localize novel instances of the regions on test datasets. We process every pixel avoiding the removal of potentially useful regions. The segmentation obtained by texton-based approach is compared against the well studied level-set approach in image segmentation community. There has been significant amount of research in computer vision literature on level-sets (cf. [144]) and applications of level-sets in medical imagery [145, 146]. The advantages of the variational level-set approach [147] incorporated in this work are: (a) the speed of curve evolution can be increased by using a large time step for numerically solving the evolution PDE, (b) the Level Set function (LSF) can be initialized with general functions that are efficient to construct and practically more viable compared with the signed distance function, and (c) the LSF is computationally efficient since it can be implemented using finite difference methods.

Binary classification of mitochondria in the cytoplasm is executed by exploiting texture features. We investigate region features based on a histogram of pixel textons. Pixel-wise classification is performed by histogram matching using a *nearest neighbor* classifier and chi-

squared statistic as a distance measure. Segmentation results demonstrate that the proposed approach using minimal training data performs close to semi-automatic carried out using variational level-set method and manual segmentation by an experienced user. Our segmentation algorithm is computationally efficient on large datasets (each tomogram is ~1-2GB), has minimal user intervention and achieves high classification accuracy as shown by experimental results. We then investigate quantitative analysis such as volume of cytoplasm occupied by mitochondria, difference between surface area of inner and outer membranes and mean mitochondrial width which are indicative quantities of segregating a cancerous cell from a normal one. To test the accuracy of our approach the quantities are compared against manually computed counterparts. In the end, we test the robustness of texture-based segmentation on tomograms of Liposomal Doxorubicin formulations (Doxil), an anticancer drug, imaged at cryogenic temperatures and study the radii and volume distribution which is of great interest in caner treatment.

5.1 Denoising Method

Denoising algorithms based on nonlinear-diffusion [96, 107, 110] and total variation methods [143] preserve sharpness and location of the edges, and in some cases enhance them. But, significant small details and textured regions of the image are disregarded. Since, our intent is to automatically segment out the mitochondria structures in MNT-1 cells (melanoma cells) that have a distinct texture, a variational denoising algorithm that preserves both the structure and texture of the image would aid in the process of automatic segmentation. The textured regions are characterized by high local variance of the residual image, I_R, where $I_R = I_0 - I$ [142]. Here, I_0 is the noisy image, and I is the filtered image. To preserve the textured regions, the filtering levels over these regions should be reduced. The denoising algorithm is based on the variational filtering using an adaptive fidelity term that applies different levels of denoising in different regions [142]. As pointed out by Gilboa *et al.*, this procedure preserves the textures and smooth regions are better denoised in terms of signal-to-noise ratio.

Filtering techniques based on gradient dependent energy functional, including non-liner diffusion [96, 107, 110] and total variation [143] noise removal techniques change images

towards piecewise constant functions. PDE [96, 107, 110] based methods have shown in impressive denoising performance on non-textured images. Image decomposition by PDE and wavelet based methods [97, 99, 101] essentially consists of three components namely structure, texture and noise are well suited for denoising natural images. These approaches dispose of the oscillatory noise while preserving or enhancing the edges. Here, noise if often regarded as spatially invariant (e.g. white Gaussian noise), but textures and structures are not homogenous throughout the image in terms of variance. This results in over-smoothing of textures in the image. The denoising algorithm used here is based on [142] that computes an adaptive spatially varying fidelity term based on the oscillatory part of the signal. This controls the extent of denoising over the image regions based on the content. The model in [142] consists of three components, the cartoon approximation, I_C, the non-cartoon component I_{NC} and additive noise I_n. The non-cartoon part consists of textures, thin lines, small-scale details etc. Therefore the noisy image can be expressed as, $I_0 = I_C + I_{NC} + I_n$ and, the residual image, $I_R = I_0 - I = \tilde{I}_{NC} + \tilde{I}_n$, where tilda upper script denotes approximation by the diffusion process. Assuming that noise is uncorrelated with the signal the variance of the residual image is the sum of the local variances of the non-cartoon component and the noise respectively. Hence, the textured regions are characterized by high local variance of the residual image and hence to retain the textured part, the filtering level needs to be reduced over these high variance regions. This also results in stronger filtering of smooth regions.

5.2 Semi-Automatic Segmentation Procedure Based on Level Set

Once we have the filtered slices from the tomogram, we implement the variational level-sets segmentation method [144]. The segmentation algorithm is based on the variational formulation of the level-set evolution without re-initialization [147]. This algorithm forces the level-set function to be close to a signed distance function, thereby eliminating the need of re-initialization procedure and speeding up the segmentation process. The re-initialization process of the level-sets to a signed distance function maintains stable curve evolution [144]. The variational energy function consists of an internal energy term that penalizes the deviation of the level-set

function from the signed distance function and an external energy drives the zero level-set to the object boundaries. Since during evolution, the internal energy keeps the level-set function close to a signed distance function, the re-initialization procedure is eliminated. The total energy function is

$$\xi(\phi) = \mu P(\phi) + \xi_{g,\lambda,\nu}(\phi), \tag{75}$$

where $P(\phi) = \int_{\Omega} \frac{1}{2}\left(|\nabla\phi| - 1\right)^2 dxdy$, $\xi_{g,\lambda,\nu}(\phi) = \lambda L_g(\phi) + \nu A_g(\phi)$, and $\phi = \phi(t,x,y)$, which is a function of the pixel location, *(x, y)*, and time *t*, is the level-set function.

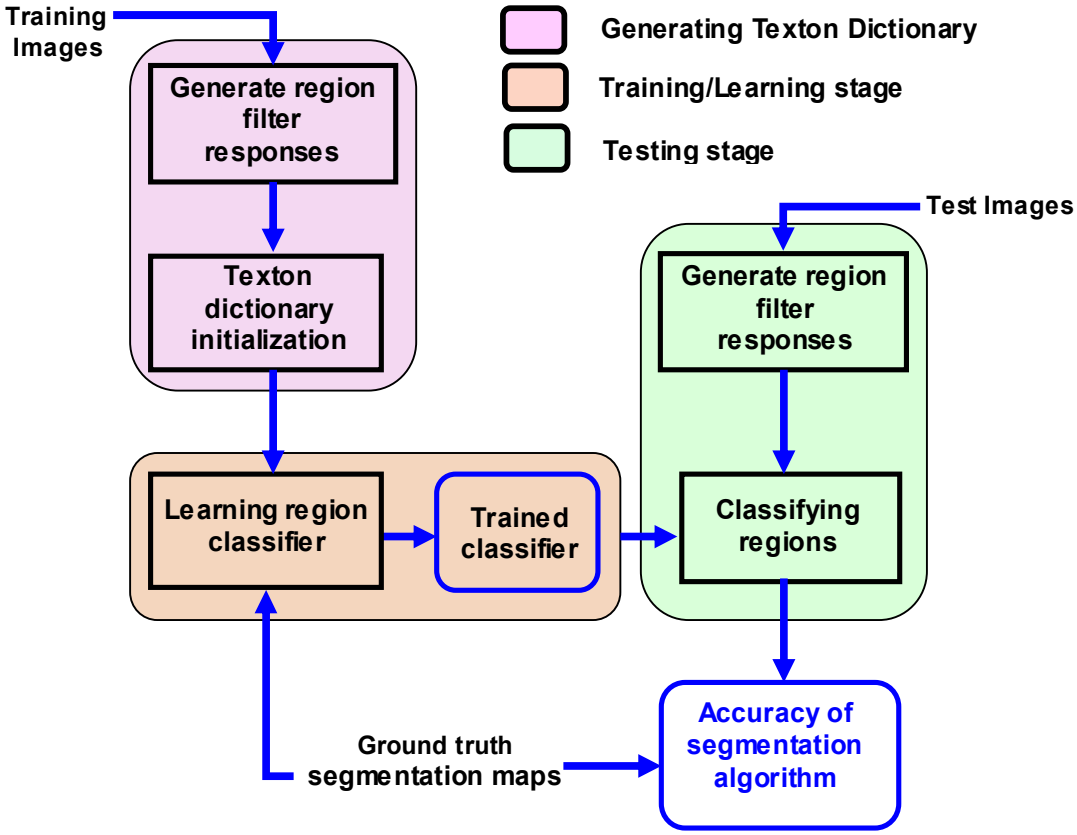


Figure 41. Texture based segmentation procedure.

$P(\phi)$ is used to characterize how close the level-set function, ϕ is to a signed distance function in $\Omega \subset \Re^2$, and $\mu > 0$ controls the effect of penalizing the deviation of ϕ from the signed distance function. In Equation (75), g is the edge indicator function, $g = \frac{1}{1 + |\nabla G_\sigma * I|^2}$, I being the input

image, and G_σ is the Gaussian kernel with standard deviation σ. $\lambda > 0$ and ν are constants. $L_g(\phi)$ is energy function that computes the length of the zero level curve of ϕ, and $A_g(\phi)$ speeds up the curve evolution. The value of ν is positive or negative depending on whether the initial contours are placed outside or inside the object. For stable level set evolution, $\tau\mu < 0.25$ [147], where τ is the time-step. The parameters such as σ (standard deviation of the Gaussian), λ, μ and ν needs to be varied for optimal segmentation for each tomogram due to low contrast and signal-to-noise ratio (SNR), appearance, geometry and viewpoint variation. Also, as a priori information, we manually provide annotated maps to separate out mitochondria from the cytoplasm. Therefore, this method is locally optimal, and hence cannot generalize well for new tomograms. This is a serious bottleneck for high-throughput data mining and quantitative analysis. Hence, we resort to machine learning approaches, where the manual intervention can be limited only to the training process

5.3 Automated Segmentation

The block diagram of the segmentation procedure is shown in Figure 41. The pixel block-based classification is divided into three stages namely generating texton dictionary, training/learning stage and testing stage. Once the segmentation maps are obtained the accuracy of the classifier is compared against the manually segmented ground truths. A subset of training images is convolved with a filter-bank to generate filter responses. The texton dictionary is created by collecting the filter bank responses after vector quantizing them, after which they are clustered using *K-means* approach, and these clusters aggregated into a texton dictionary. These clusters form the basis-set for describing the texture images both in the training as well as in the testing stages. We use a filter-bank that is similar to that of Winn [117] which consists of 3 Gaussians, 4 Laplacian of Gaussians (LoG), and 4 first-order derivatives of Gaussians. The Gaussian and the LoG filters (isotropic) are rotationally symmetric. The RGB image is converted to the CIE Lab space since they are more robust and the Gaussian kernels ($\sigma = 1, 2, 4$) are applied to each of these channels resulting in a total of 9 filter responses. The LoG kernel ($\sigma = 1, 2, 4, 8$)

and the four derivative of Gaussian kernel ($\sigma = 2, 4$, two each in x and y directions) are applied to the L channel alone. The derivative-of-Gaussian filter provides good features for anisotropic textures (edges at different orientations) as they are not rotationally invariant. Since our set consists of both isotropic and anisotropic filters, they are expected to generate robust texture features. So, each pixel is associated with a 17-dimensional feature vector, and we keep this to a minimum set as opposed to [119, 129, 131] to limit the extra processing and computational cost and also to avoid the clustering issue due to high dimensionality.

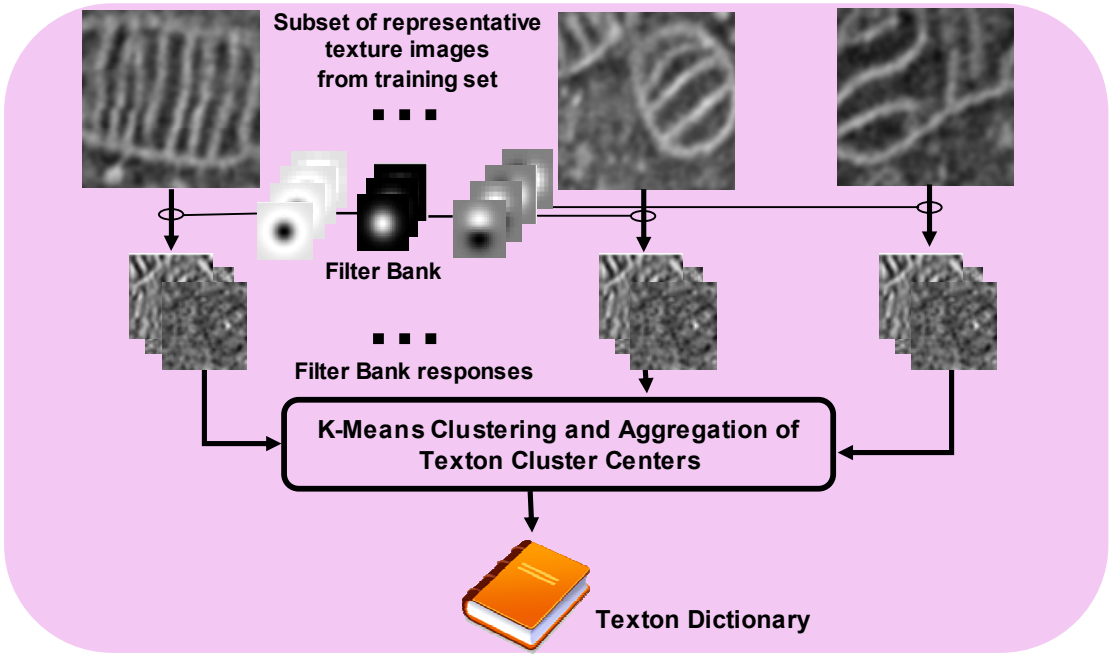


Figure 42. Texton dictionary generation.

In our implementation, we selected 38 randomly chosen images from the training set such that the texton dictionary created spans the texture classes. Our approach differs from the prior approaches [117, 119, 129] in the sense that we use "minimal training set " which consists of only 256x256 sized sub-images that spans all the classes to create the texton dictionary. Examples of training images are shown in Figure 42. Since it is computationally expensive to use features associated with each pixel in a sub-image, we randomly permute and pick first L pixels, and we vary L between 1000 to 2500 pixels. The filter responses of these 38 sub-images are aggregated and 50 cluster (25 clusters/texture class) centers are (textons) learnt. The textons that are learnt from each of these textures form a single dictionary.

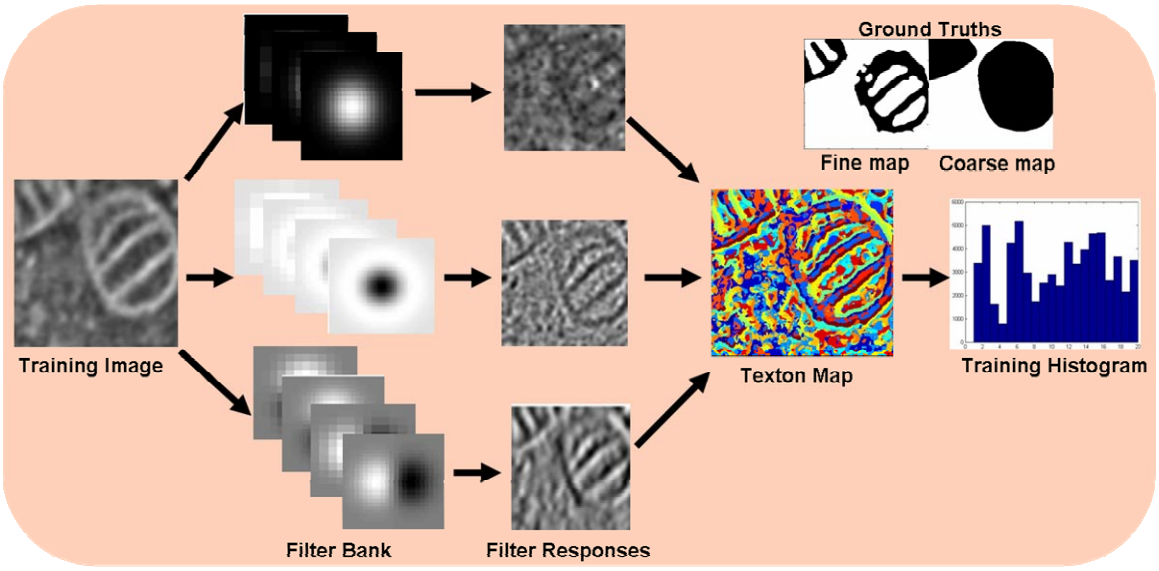


Figure 43. Training and Learning stage.

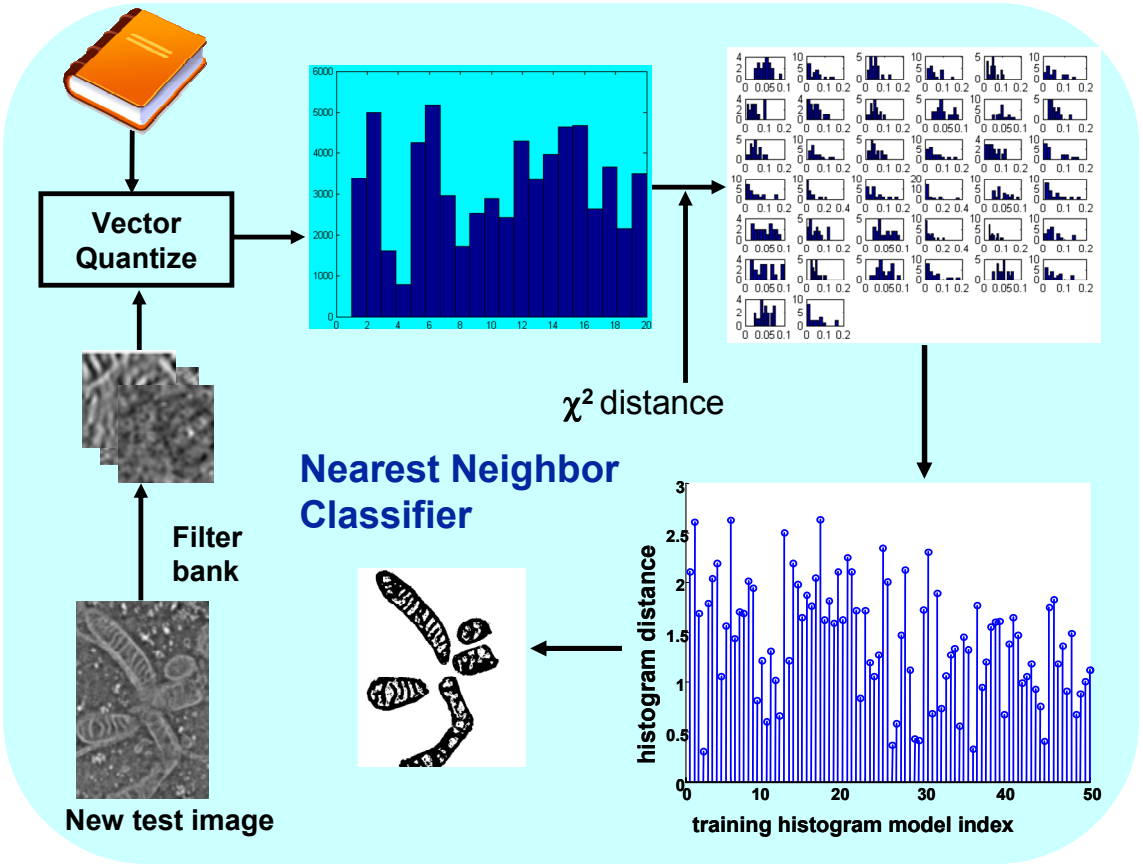


Figure 44. Testing stage

After creating the texton dictionary, the next stage is to learn models for each of the texture classes. This is shown in Figure 43. Here, each of the training sub-images are convolved with the filter bank and labels are assigned to each filter response with the texton that lies closest to it in the filter response space using Euclidean distance measure.

The training histogram is nothing but the frequency with which each texton occurs in the labeling. We use manually labeled images that contain both segmentation maps and rough annotations of regions as shown in Figure 43, to learn the model corresponding to each training image. The histograms are normalized since it achieves robustness to scale variations. At this stage, multiple training histograms are used to characterize each texture class. In the testing stage, we chose an independent tomogram of MNT-1 cell compared to the training set. A similar procedure is followed for each image slice in the tomogram as in the training stage. There are two reasons for using 2D segmentation compared to 3D: first, the slicing of the specimen by ion beam in the increments of ~20nm reduces the dependence between individual slices and second, the computational savings is considerable compared to 3D segmentation. Once the histograms corresponding to the test image are constructed using block processing (we chose an *MxM* block, where *M=3*), these histograms are compared using chi-square distance statistic with the learnt training histograms. A *nearest neighbor* classifier is employed, and the block of pixels is assigned the class label that it is closest to using the distance metric. The procedure is summarized in Figure 44.

5.4 Simulation Results

We consider two datasets of mitochondria in MNT-1 cells imaged using an ion-abrasion scanning electron microscope. Both the datasets are of low contrast and SNR; hence, we use texture preserving denoising [142] to enhance the texture part of the tomogram. Mitochondria in MNT-1 cells have a unique band-like texture that can be exploited in the process of automatic segmentation. Before filtering, each slice was filtered using an adaptive notch filter to remove the vertical lines that are caused due to milling during the imaging procedure. We applied the variational texture-preserving filtering procedure for each slice in the tomogram with

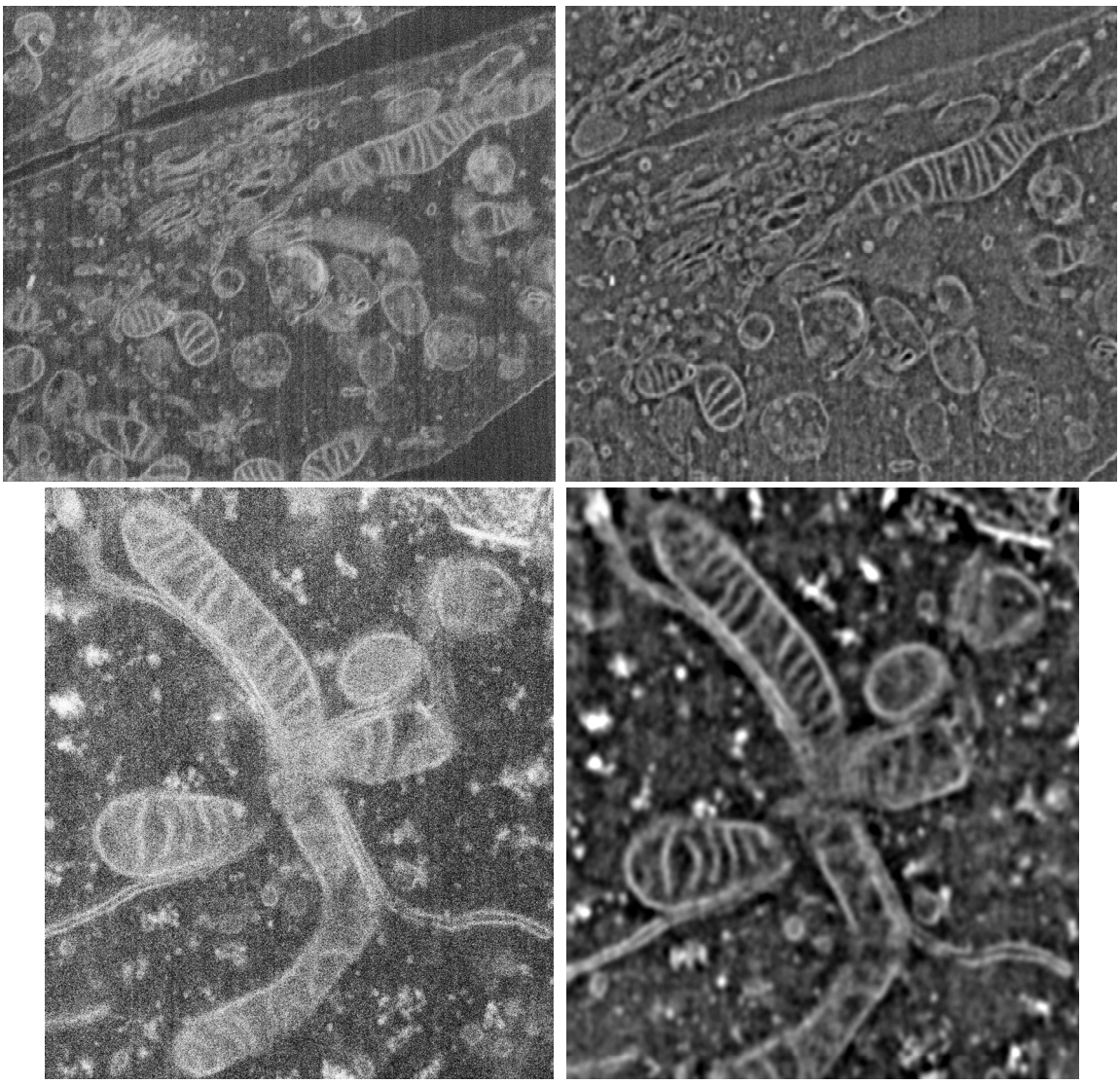


Figure 45. Denoising using variational texture-preserving filter

the same set of parameters. The results of denoising on both the training and test datasets are depicted in Figure 45 for 10 iterations. Next, we implemented variational level-set segmentation algorithm based on [147] on each slice of the tomogram. This is a semi-automatic technique since we need to manually provide rough annotated maps to separate out mitochondria from the cytoplasm. In our experiments, we used a positive value for ν, $\mu = 0.04$, $\tau = 5$, $\lambda = 5$, and $\sigma = 1.5$. We also varied the annotated maps between (1-11) which is about (1.4%-15%) of the total number of slices as prior information. There were about 72 and 68 slices respectively in the two datasets. So we only needed 15% (~11 maps) of roughly annotated maps to obtain

segmentation close to or better than that obtained manually by an experienced user. After separating out the mitochondria-like structures, we used the 2D maps to generate the 3D volume.

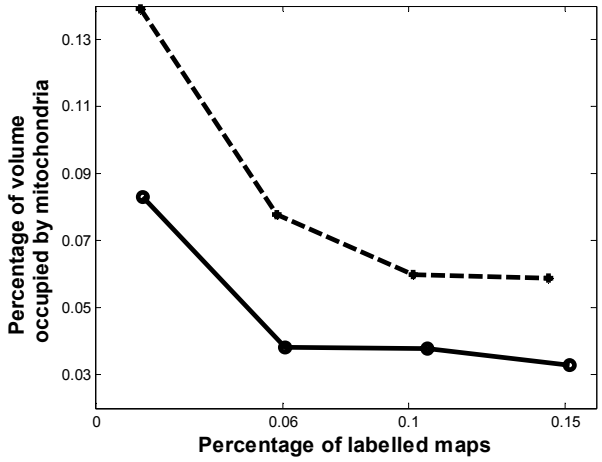


Figure 46. Percentage of the volume occupied (segmented using variational level-sets) by the mitochondria in the cytoplasm as a function of the labeled maps.

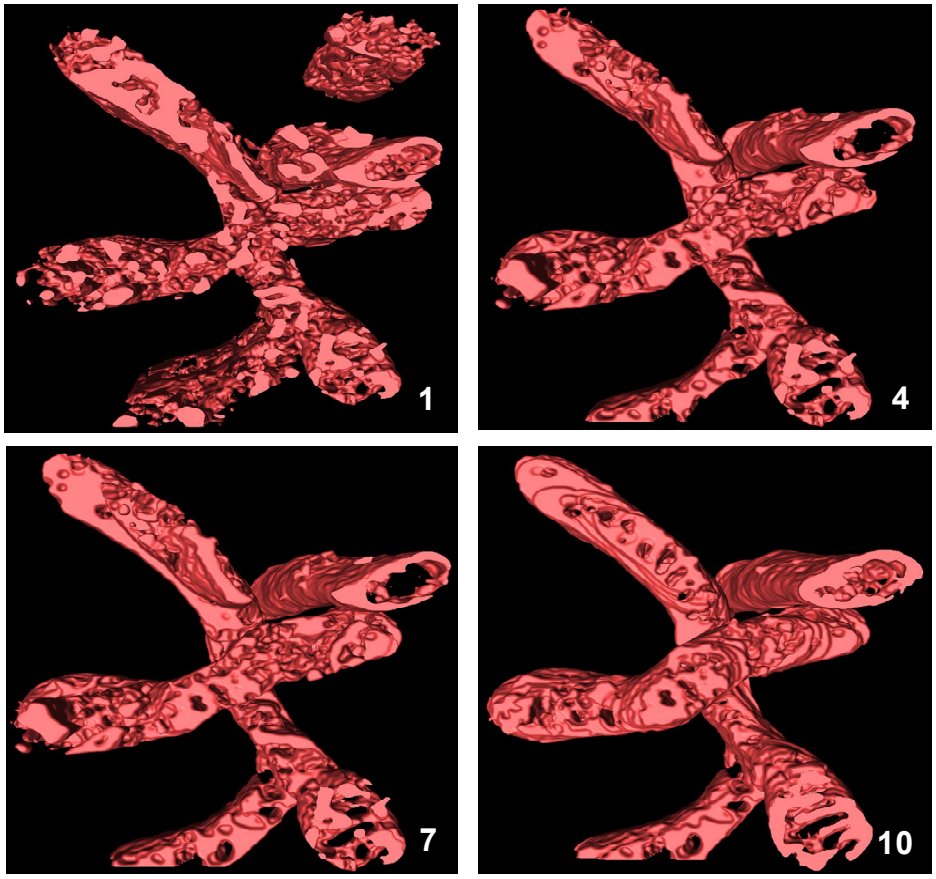


Figure 47. Segmented volume using variational level-set method. The numbers indicate the number of labeled maps that were used.

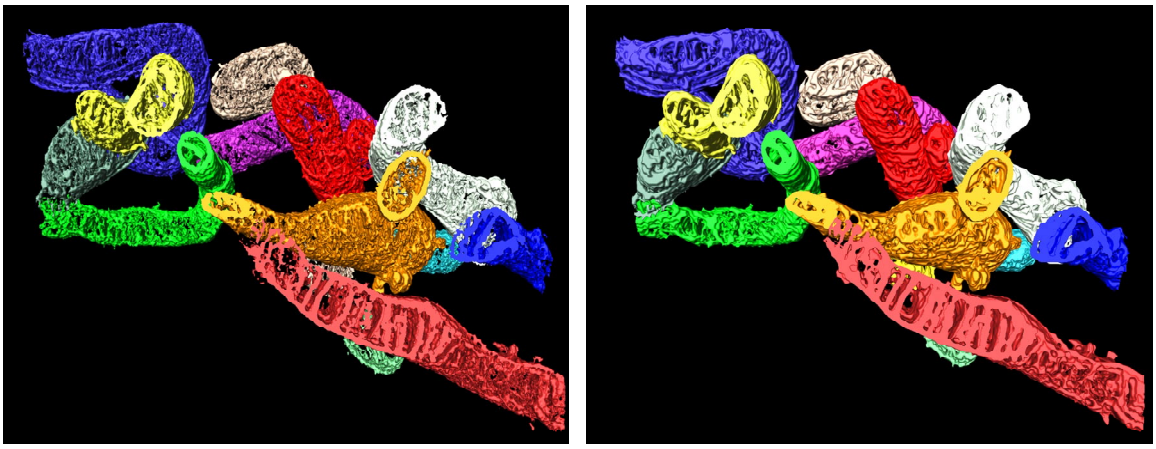


Figure 48. Comparison of the segmented volume using variational level-set method (right) against manual approach (left).

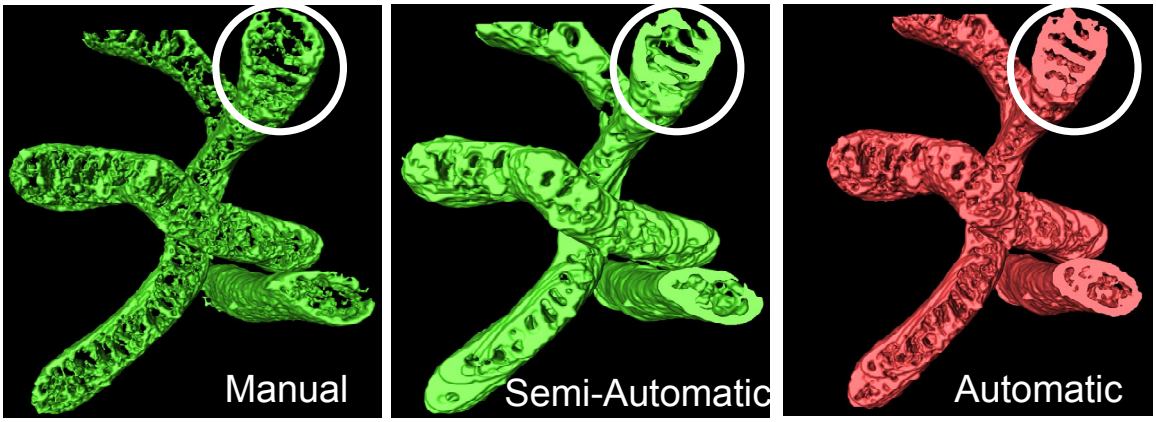


Figure 49. Comparison of the manual, semi-automatic (using variational level-sets) and automatic texton-based approaches. The circle indicates the improvement in segmentation progressively from left to right.

We also measured the percentage of volume occupied by the mitochondria compared to the cytoplasm as we increased the percentage of annotated maps, as shown in Figure 46. Their respective volumes are depicted in Figure 47. We found that the actual percentage was between (4%-7%) of the total volume and this was confirmed using the volume obtained from manual segmentation. The manual and semi-automated segmentations are used as benchmarks for comparing the segmentation volumes obtained using texton-based approach. Figure 48 shows the

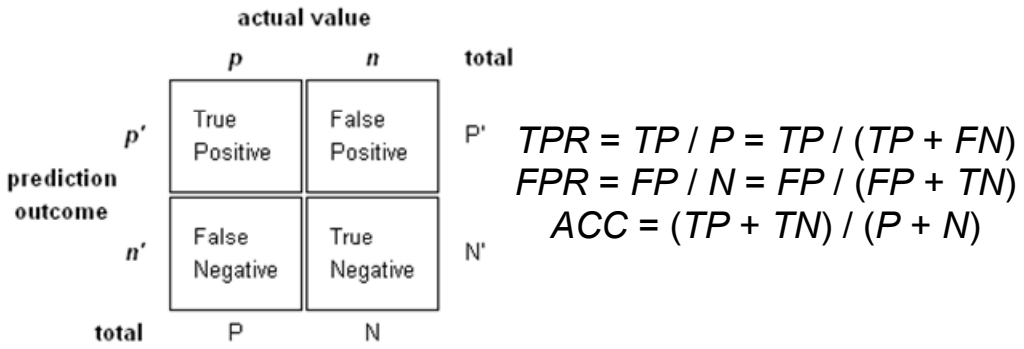


Figure 50. ROC curve parameter estimation.

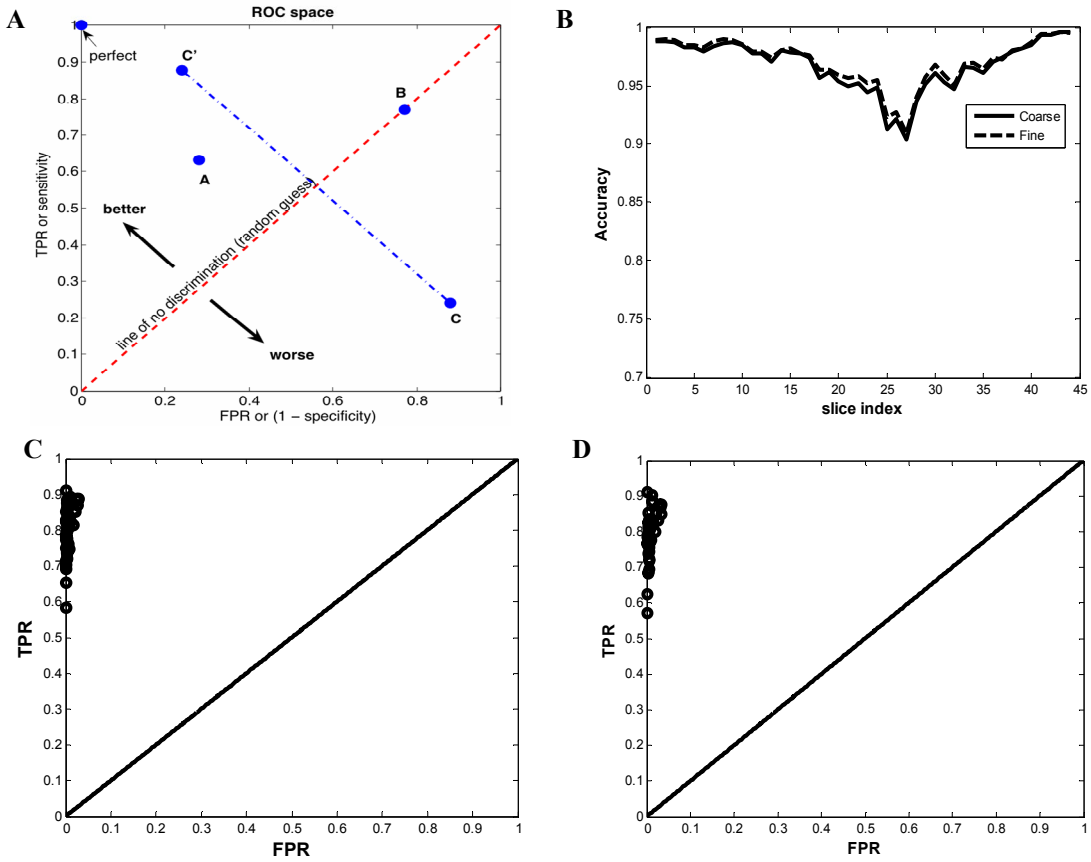


Figure 51. (A) Receiver operating characteristic (ROC) curves. (B) Accuracy plot of the nearest neighbor classifier using both the segmented maps and the rough annotation in the training stage. (C & D) ROC plot for the nearest neighbor classifier using both the segmented maps and the rough annotation in the training stage.

comparison of the respective volumes using the manual and semi-automated procedure on the dataset that was used for creating the texton dictionary. As mentioned earlier, in the texton-based approach we trained the classifier on one dataset of mitochondria and tested on the other. Figure 49 compares the volumes obtained manually by an experienced user, the semi-automatic segmentation procedure based on variational level-sets that takes into consideration prior information using the labeled maps, and the automatic texture-based segmentation approach. The "circle" in each of the figures demonstrates the improvement of segmentation progressively from manual to automatic. The reason is that both the level-set and the texture-based approaches provide a maximal connected surface as a segmented map. The receiver operating characteristic (ROC) curve, which is the plot of the true positive rate (TPR) vs. the false positive rate (FPR), is shown in Figure 51(A). The accuracy is obtained as shown in Figure 50.

The ROC curves for the texture-based approach using both the segmented maps and rough annotation is depicted in the Figure 51(C) & (D). The accuracy plot for the same is shown in Figure 51(B), where the accuracy minimum and maximum was between (91-99) percent. Next, we compare the volume and the segmented maps obtained using texton-based approach against that of variational level-set method in top and bottom rows of Figure 52 respectively. As seen from the results the segmented maps obtained using texton-based approach is better connected than the manual and the semi-automatic counterpart and their respective volumes are also comparable. It is also interesting to note that the segmentation obtained by using rough annotated maps, and with the segmented maps in the training stage are also close to each other which indicate the robustness of the texton-based approach. This is also verified by the accuracy and the receiver operating characteristic (ROC) curve plots computed for each slice depicted in Figure 51. Results demonstrate that using minimal training data, the texton based approach performs close to the segmentation results obtained by the semi-automatic variational level-set method.

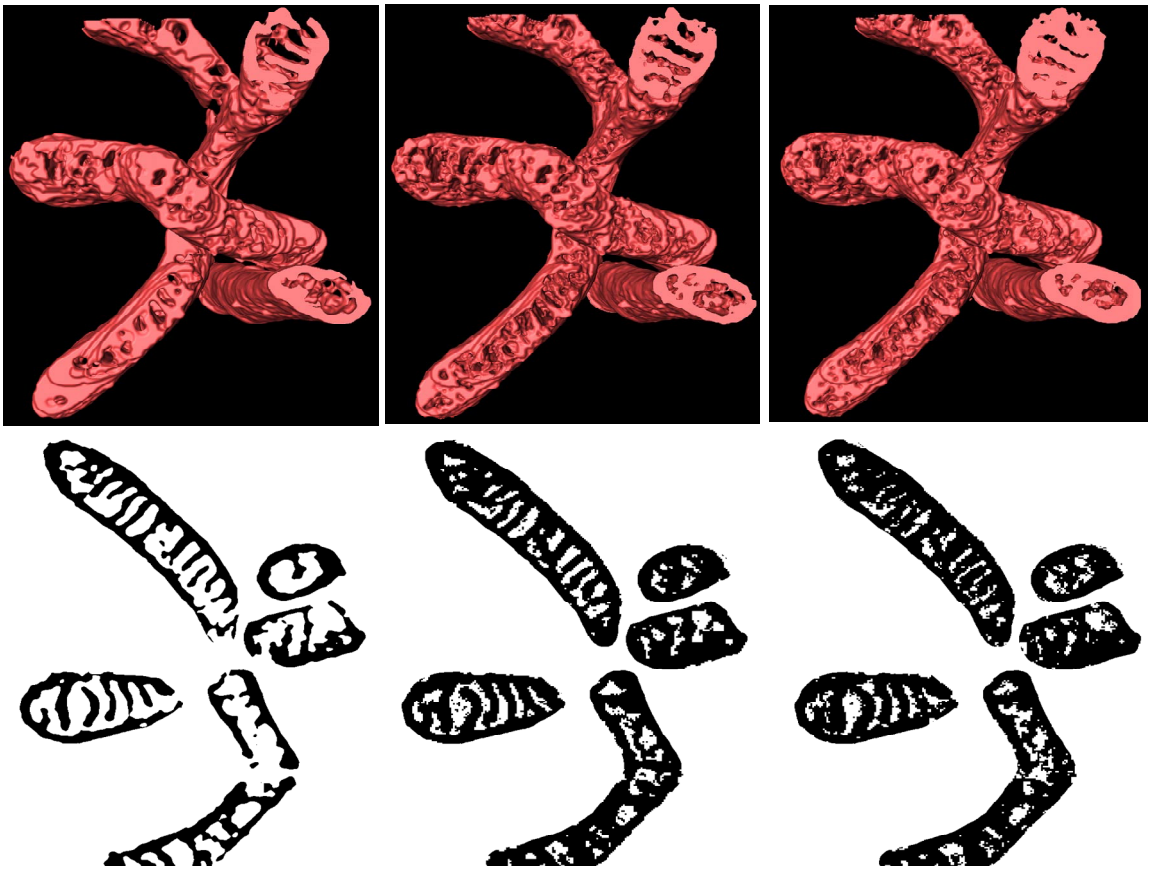


Figure 52. Comparison of the volume (Top Row) and segmented maps (Bottom Row) of the semi-automatic (using variational level-sets) and automatic texton-based approaches (using segmented maps and rough annotation).

We then investigate quantitative analysis such as the surface area of cytoplasm occupied by mitochondria in each slice of the tomogram, the difference between surface area of inner and outer membranes, and the mean mitochondrial width which segregate cancerous cells from a normal ones. The plot of the surface area of cytoplasm occupied by mitochondria per slice comparison using the texton-based and level-set methods. The surface area comparison is demonstrated in Figure 53. Next, we obtain the average difference between the inner and outer membranes in respectively. We hypothesize an approximate expression for the membrane differences as

$$l_{inner} \approx l_{outer} + 2\sum_{i} d_i \,, \tag{76}$$

where l_{inner} and l_{outer} the average length of the inner and outer membranes respectively whereas d is the mitochondrial width. The mean and standard deviation of the average difference is consistent for both the datasets and the maximum difference was found to be approximately two fold. Table 4 and Table 5 depict the mean width comparison for both the datasets respectively. The mean widths obtained by the texton-based method are comparable to those obtained using manual and level-set approaches.

5.5 Quantitative Analysis

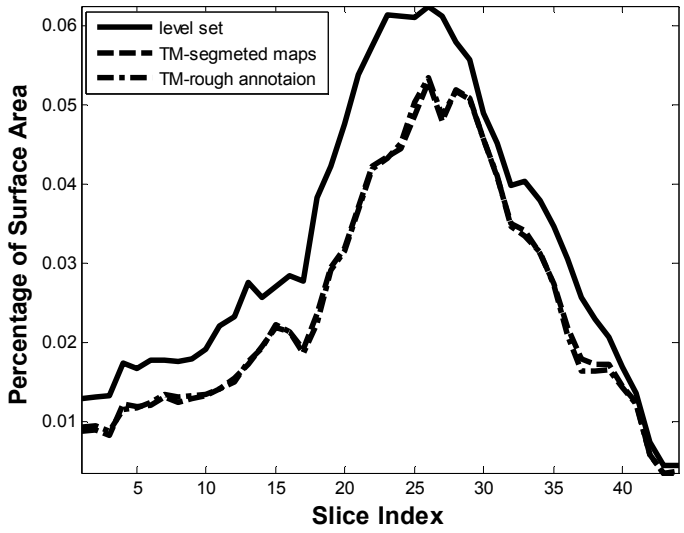


Figure 53. Comparison of the surface area occupied by mitochondria in the cytoplasm as a function of the slice index.

Table 4. Mean widths comparison for dataset-1

Dataset -1	Mean Width in nm
Manual	391.2574
Level Sets	392.5641

Table 5. Mean widths comparison for dataset-2

Dataset – 2	Mean Width in nm
Manual	329.9011
Level Sets	339.5229
Texton Method : Segmentation Maps	343.9592
Texton Method : Rough Annotation	373.7955

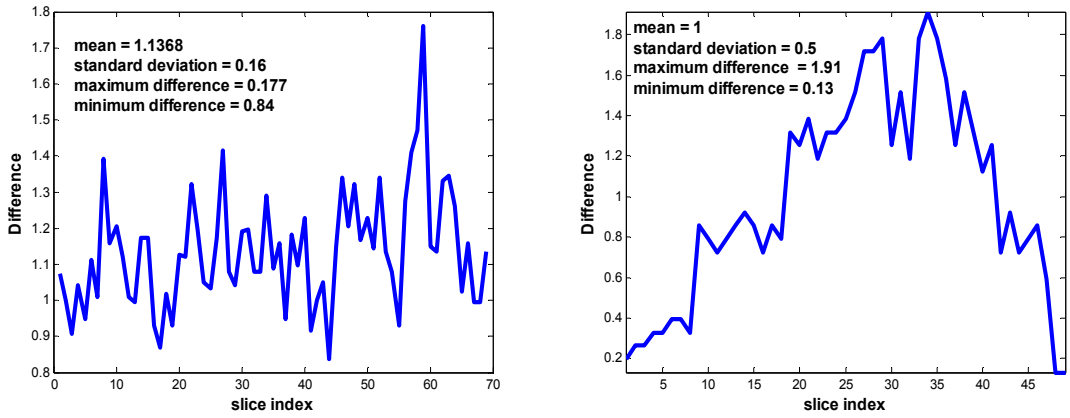


Figure 54. Difference between the inner and outer membrane lengths.

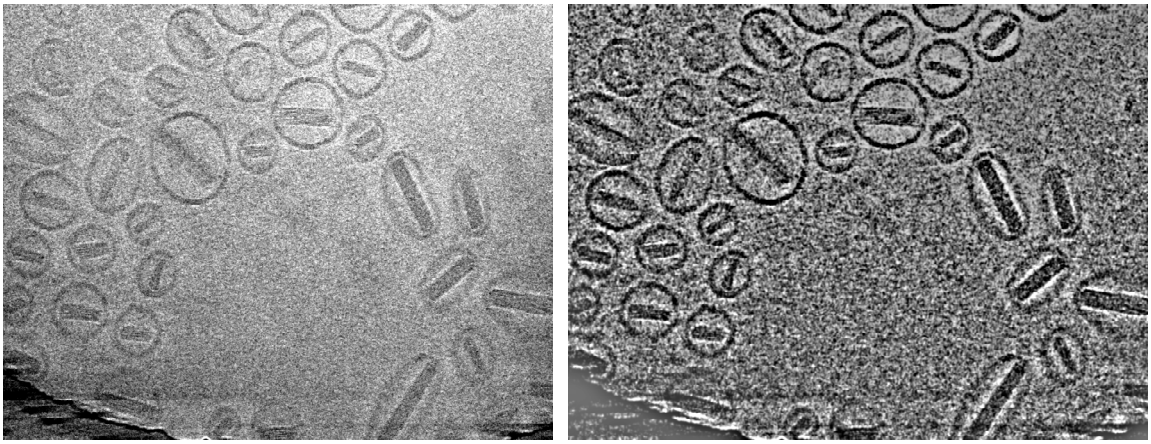


Figure 55. Denoising using non-linear anisotropic diffusion

Furthermore, we test the robustness of our texture-based segmentation approach on tomograms of Liposomal Doxorubicin formulations (Doxil), an anticancer drug, imaged at cryogenic temperatures and study the radii and volume distribution which is of great interest in caner treatment. By targeting it to various cells in the body based on its radii distribution, this nano-drug achieves specific diagnosis. The study of the radii distribution may aid in reducing the side effects caused by this drug in cancer treatment. We used a non-linear anisotropic diffusion denoising method [96] for these datasets as a preprocessing step. The result is shown in Figure 55. As before, we used one dataset in the training stage and tested our approach on an independent dataset used in training. We also used a semi-automated segmentation technique

based on variational level-set approach. The difference compared to the mitochondrial segmentation is that we used only three rough annotated maps in this case as prior information. The volume is shown in Figure 56. The result of the automatic segmentation using texture-based approach is demonstrated in Figure 57 .

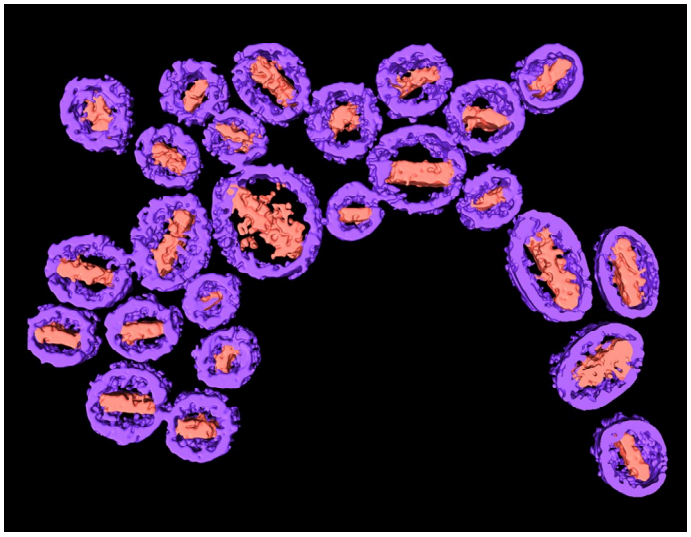


Figure 56. Semi-automated segmentation using variational level-sets

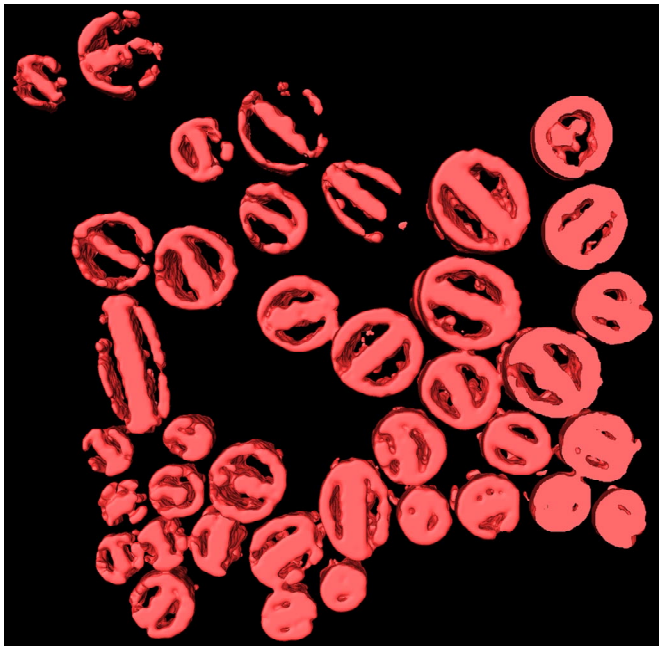


Figure 57. Automated segmentation using texture-based approach

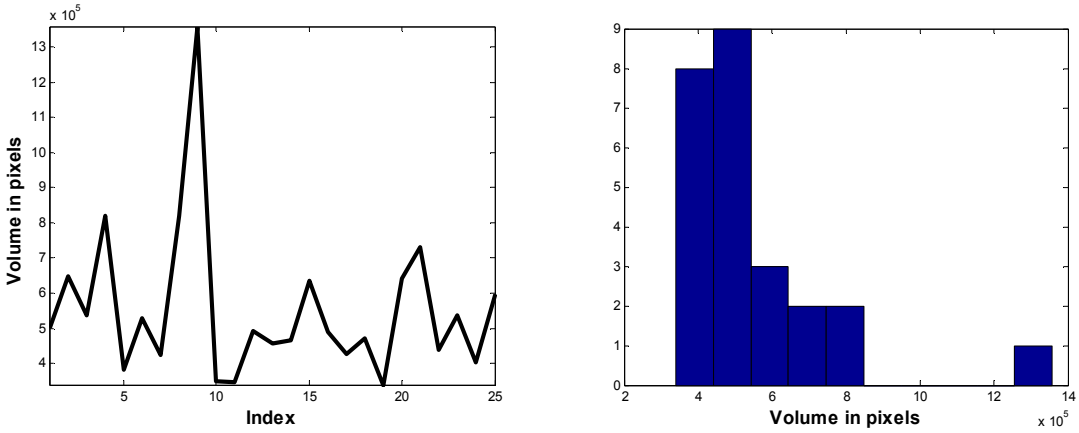


Figure 58. Volume distribution of Liposomal Doxorubicin formulations

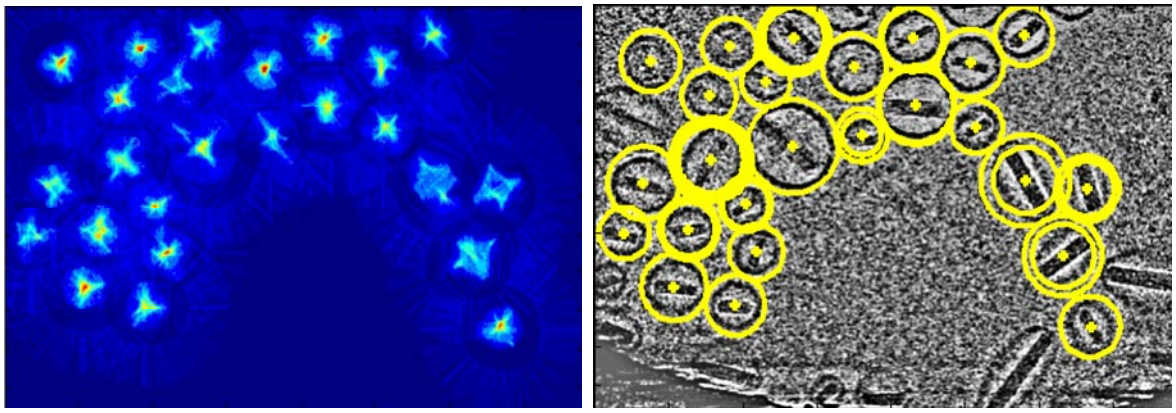


Figure 59. Radii estimation using Hough transform.

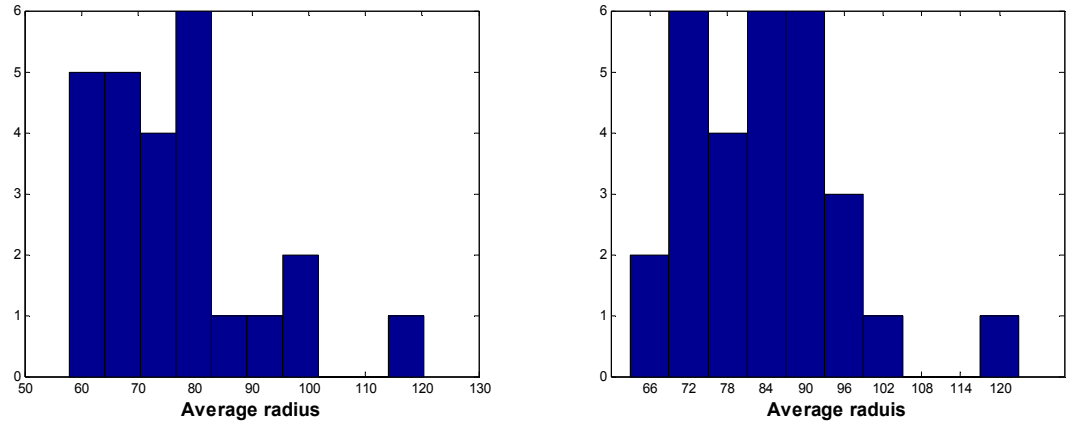


Figure 60. Comparison of the radii distribution using manual (left) and Hough transform (right).

The volume of each of the Liposomal Doxorubicin formulations and their distribution is shown in Figure 58. We also computed radius of individual doxils using the Hough transform as depicted in Figure 59. In some cases, where the shape was not perfectly circular, we used the average of the major and minor axes. The comparison of the radii distribution obtained using manual and Hough transform methods are depicted in Figure 60, and the *p-value* of chi-square test was calculated based on the null hypothesis that both histograms are sampled from identical distributions. Here, the *p-value*, which lies between 0 and 1, is the lowest significant level at which the null hypothesis can be rejected if $p<0.1$. The *p-value* of the chi-square test was 0.8647, indicating that the null hypothesis is plausible.

5.6 Summary

This chapter investigated automatic techniques for segmentation and quantitative analysis of mitochondria in MNT-1 cells imaged using an ion-abrasion scanning electron microscope [148], and tomograms of Liposomal Doxorubicin formulations (Doxil), an anticancer nano-drug, imaged at cryogenic temperatures. As a preprocessing step, the tomograms were denoised by a two-dimension texture-preserving filter that incorporates a spatially varying fidelity term that controls the denoising of local textured regions.

A machine learning approach was formulated that exploits texture features, and joint image block-wise classification and segmentation was performed by histogram matching using a *nearest neighbor* classifier and chi-squared statistic as a distance measure. Given manually labeled images that contain both segmentation maps and rough annotations of regions, the algorithm learns models that can localize novel instances of the regions on diverse test datasets. Segmentation results demonstrate that the proposed approach, using minimal training data, performs close to or in some cases better than semi-automatic segmentation carried out using variational level-set method and manual segmentation by an experienced user.

Quantitative analyses of mitochondria in MNT-1 cells investigated in this work are indicative of early detection of melanoma, the highest cause for deaths in skin related cancer. The mean mitochondrial width was ~330nm, and the difference between the surface area of the inner and outer membranes was ~2 fold. All the automatically computed quantities were compared

against manually obtained counterparts and found to be within the acceptable error range. In the end, to test the robustness of the approach, automated segmentation and quantitative analysis of radii distribution of doxils were investigated, which may be helpful in reducing the side effects caused by this nano-drug in cancer treatment.

CHAPTER 6

CONCLUSION

This thesis dealt with the application of information theory and machine learning to solve inference problems in large networks, and developed automated methods to mine relevant byte-sized information from gigabyte-sized tomograms. We now summarize the main results in this thesis.

6.1 Main Contributions

- We considered theoretical and practical management resources needed to reliably diagnose congested or faulty network elements, and provided fundamental limits on the relationship between the number of probe packets, the size of the network, and the performance of diagnosing such network elements. We formulated the scalability of measurement-based network monitoring as a machine learning problem using probabilistic graphical models that characterize the spatial and statistical dependence of path-based probe measurements.
- We have applied variational inference and source-coding techniques to failure diagnosis to derive a lower bound on the average number of needed measurements with respect to the size of the network in the presence of noisy probe outcomes. We have shown that if the number of observations was less than the lower bound, we are guaranteed not to achieve a zero diagnosis error asymptotically, where the error was quantified in terms of the most probable explanation *(MPE)* error and the bit-error rate *(BER)*. We have derived explicit expressions for the scaling constant that multiplies the growth function, and these together provided at most linear growth rate with respect to the size of the network. Then, we examined the growth rate of the average number of measurements for non-scalable scenarios, and showed that it increases at a polynomial rate with respect to the network size. Scalability results were verified using simulations on a synthetic network. We studied optimal probe length strategy in the presence of noise. By mapping the noisy network probing scenario to a binary

symmetric channel in source coding, we developed a lower bound on the average number of probe measurements in terms of the entropy of the prior probability of link failure and the noise parameters. We then compared the lower bound obtained using the variational method against the entropy lower bound.

- We considered the congestion localization problem, where we derived theoretical bounds on the growth rate of the number of required measurements by relating the congestion localization in networks to the problem of decoding linear error correcting codes (ECC) over a binary symmetric channel (BSC) in coding theory. We evaluated the maximum fraction of the congested nodes that can be localized with zero detection error by fixing the number of measurements, and then we examined the growth rate of the number of needed measurements as the percentage of congested nodes increased. As in the link-failure diagnosis case, we provided expression for the scaling constant. We devised a methodology to accurately localize nodes that are congested in real time using a bit-flipping mechanism, which was implemented in real-time by a belief propagation algorithm. This was similar to decoding irregular low-density parity check (LDPC) codes. Simulations were performed on networks of various sizes and under noiseless and noisy probing scenarios and the performance was compared with the fundamental limits. The monitoring approach required just one bit of payload in the probe packet for congestion localization, and hence the approach was shown to be practically realizable on networks where priority queuing is implemented. Thus, the approach is computationally efficient and can be implemented without overburdening the core routers.
- In the next part of the dissertation, we focused on problems in electron tomography, wherein for analysis and feature extraction of biological tomograms, we used datasets of HIV-1, recorded at room temperature, and reconstructed using SIRT-based approaches. We implemented dual-axis data collection in this work, since it has less missing information in the tomogram, and we used SIRT because of the improved SNR in the reconstructed tomograms as compared to conventional weighted back projection (WBP) methods. The other tomograms that we have used for analysis and feature extraction were of *Bdellovibrio*, a small bacterium, recorded at cryogenic temperatures, and reconstructed using a WBP-based

approach. We reported on the relative merits of a variety of image/transform domain denoising algorithms on the detection, clustering, and spatial distribution of spikes in SIV/HIV viruses, and the automated segmentation of ribosomal complexes in the *Bdellovibrio* bacterium. Using quantitative measures such as single-image SNR estimation, Fourier ring correlation, and the KL-distance based goodness-of-fit techniques and without assuming any prior noise model we identified an "optimal" denoising strategy for tomograms imaged at room and cryogenic temperatures. We found that the phase preserving algorithm (in the case of room temperature tomograms) and the nonlinear anisotropic diffusion (NAD) algorithm (in the case of tomograms imaged at cryogenic temperatures) significantly improved the SNR, while retaining the same overall spatial architectural information in the tomograms.

- Following the denoising, molecular information such as the average volume of individual HIV viral spikes was obtained from denoised tomograms of HIV-1 infected macrophages, which was validated against a manual approach. Spatial information such as ribosome distribution was obtained automatically from denoised tomograms of *Bdellovibrio* using a template matching-based feature extraction method that considers intensity and mean-squared error between the template and the ribosome in the tomogram. The results closely matched those obtained using semi-automated approaches.
- A machine learning approach was formulated for automatic segmentation of mitochondria in MNT-1 cells imaged using an ion-abrasion scanning electron microscope, and tomograms of Liposomal Doxorubicin formulations (Doxil), an anticancer nano-drug, imaged at cryogenic temperatures. Segmentation results demonstrate that the proposed approach, using minimal training data, performs close to or in some cases better than semi-automatic segmentation carried out using variational level-set method and manual segmentation by an experienced user.
- Quantitative analyses of mitochondria in MNT-1 cells investigated may facilitate early detection of melanoma, the highest cause of deaths in skin related cancer. The radii distribution of doxils was investigated, which may be helpful in reducing the side effects caused by this nano-drug in cancer treatment.

6.2 Future Work

- In principle, if we could design an parity-check matrix *(i.e., routing matrix A)* that can achieve the lower bound by reducing the cycles in the bipartite graph, e.g. as in a MPLS network using "explicit routing" [149], then design the probing paths, the number of probes can be reduced significantly. As part of further investigation, it would be interesting to assess the bounds using measurements from operational networks. Furthermore, our method can be extended to the scenario where the end-hosts form an overlay network that assumes knowledge of the underlying IP topology, and also to detect security holes and localizing attacks in networks.
- Noise in the tomogram depends on various factors such as specimen preparation, data collection schemes, and using different algorithms to align and reconstruct the data into a final 3D volume. The study of noise in tomograms using these individual factors and identifying optimal and robust denoising methods can be valuable for post processing such as segmentation, classification, and quantitative analysis.
- The texture-based segmentation presented in this thesis can be further improved using filter banks designed for specific types of tomograms, and support vector machines (SVM) and boosting techniques combined with cross-validation for classification. Also, the next step would be to apply the developed techniques to identify quantitative methods to segregate cancer from non cancer cells for early diagnosis.

APPENDIX A

Proof of theorem 1:

Since $\theta_{ij} = -\log(1-p_{ij})$ we have, $P(y_i^+ \mid x) = e^{-\theta_{i0}-\sum_{j\in\pi_i}\theta_{ij}(1-x_j)}$ and $P(y_i^- \mid x) = 1-e^{-\theta_{i0}-\sum_{j\in\pi_i}\theta_{ij}(1-x_j)}$. Now,

$$\begin{aligned} Q(X,Y) &= \max_x P(X,Y) \\ &= \max_x \left(\prod_{j=1}^{n_f} P(x_j)\right)\left(\prod_{i=1}^{m_f} P(y_i \mid x)\right) = \max_x \left(\prod_{j=1}^{n_f} P(x_j)\right)\left(\prod_{i=1}^{m_f}\left[P(y_i^+ \mid x)\right]^{y_i}\left[P(y_i^- \mid x)\right]^{1-y_i}\right) \end{aligned} \quad (77)$$

$$\begin{aligned} &\le \max_x \left(\prod_{j=1}^{n_f} P(x_j)\right)\max_x \left(\prod_{i=1}^{m_f}\left[P(y_i^+ \mid x)\right]^{y_i}\left[P(y_i^- \mid x)\right]^{1-y_i}\right) \\ &= \max_x \left(\prod_{j=1}^{n_f} P(x_j)\right)\max_x \left(\prod_{i=1}^{m_f}\left[e^{-\theta_{i0}-\sum_{j\in\pi_i}\theta_{ij}(1-x_j)}\right]^{y_i}\left[1-e^{-\theta_{i0}-\sum_{j\in\pi_i}\theta_{ij}(1-x_j)}\right]^{1-y_i}\right), \end{aligned} \quad (78)$$

where $\phi_f = \max_{j\in\{1,2,\ldots n_f\}}\left\{\max\left(\rho_{f_j}, 1-\rho_{f_j}\right)\right\}; \rho_{f_j} = P(X_j = 0)$. Let $\psi_i = \theta_{i0} + \sum_{j\in\pi_i}\theta_{ij}$ and the second term in the above equation can be written as $1-e^{-\theta_{i0}-\sum_{j\in\pi_i}\theta_{ij}} = 1-e^{-\psi_i} = e^{\log(1-e^{-\psi_i})} = e^{f(\psi_i)}$. Now, $f(\psi_i) = \log(1-e^{-\psi_i})$ is a concave (convex down) function. Hence,

$$f(\psi_i) \le \xi_i\psi_i - f^*(\xi_i) \Rightarrow e^{f(\psi_i)} \le e^{\xi_i\psi_i - f^*(\xi_i)} = e^{\xi_i(\theta_{i0}+\sum_{j\in\pi_i}\theta_{ij}) - f^*(\xi_i)}. \quad (79)$$

The conjugate function ($f^*(\xi)$) for a noisy-OR network, from [22], is

$$f^*(\xi) = -\xi\log(\xi) + (1+\xi)\log(1+\xi). \quad (80)$$

From equation (78), using the tangent plane as a bound, we have,

$$Q(X,Y) \le \phi_f{}^{n_f}\left(\prod_{i=1}^{m_f}\left[e^{-\theta_{i0}}\right]^{y_i}\left[1-e^{-\theta_{i0}-\sum_{j\in\pi_i}\theta_{ij}}\right]^{1-y_i}\right) \le \phi_f{}^{n_f}\left(\prod_{i=1}^{m_f}\left[e^{-\theta_{i0}}\right]^{y_i}\left[e^{\xi_i(\theta_{i0}+\sum_{j\in\pi_i}\theta_{ij})-f^*(\xi_i)}\right]^{1-y_i}\right). \quad (81)$$

Let $\alpha_i = \xi_i\left(\theta_{i0} + \sum_{j\in\pi_i}\theta_{ij}\right) - f^*(\xi_i)$. Therefore, $Q(X,Y) \le \phi_f{}^{n_f}\left(\prod_{i=1}^{m_f}\left[e^{-\theta_{i0}}\right]^{y_i}\left[e^{\alpha_i}\right]^{1-y_i}\right)$.

Assuming homogenous conditions, we have, $\alpha_i = \xi_i(\theta_0 + \theta r_i) - f^*(\xi_i)$, hence

$$Q(X,Y) \le \phi_f{}^{n_f}\left(\prod_{i=1}^{m_f} e^{-\theta_0 y_i} e^{\alpha_i(1-y_i)}\right) = \phi_f{}^{n_f}\left(\prod_{i=1}^{m_f} e^{-(\theta_0+\alpha_i)y_i} e^{\alpha_i}\right) = \phi_f{}^{n_f}\left(\prod_{i=1}^{m_f} e^{\beta_i y_i} e^{\alpha_i}\right), \quad (82)$$

where $\beta_i = -(\theta_0 + \alpha_i)$. The optimum variational parameter (ξ_i^{opt}) can be obtained by maximizing the exponent in (82) as $\xi_i^{opt} = \frac{e^{-(\theta_0 + \theta r_i)}}{1 - e^{-(\theta_0 + \theta r_i)}}$. Now evaluating the MPE error, we have the result.

$$1 - \sum_y \max_x P(X = x, Y = y) \geq 1 - \sum_y \phi_f{}^{n_f} \left(\prod_{i=1}^{m_f} e^{\beta_i y_i} e^{\alpha_i} \right). \tag{83}$$

APPENDIX B

Proof of Corollary 1:

Under the equal probe route length assumption, $\xi_i = \xi, \beta_i = \beta, \alpha_i = \alpha$, and we bound the *MPE error* by the performance bound as $MPE \le \delta_{f_1}$. The optimal variational parameter simplifies to $\xi_i^{opt} = \xi$. Suppose there are exactly *k* positive evidences, then (34) simplifies as

$$1 - \phi^{n_f} (e^{-\theta_0} + e^{\alpha})^{m_f} \le \delta_{f_1}, \tag{84}$$

$$\phi^{n_f} (e^{-\theta_0} + e^{\alpha})^{m_f} \ge 1 - \delta_{f_1}. \tag{85}$$

Taking logarithm on both sides, the result is obtained.

APPENDIX C

Proof of Theorem 2:

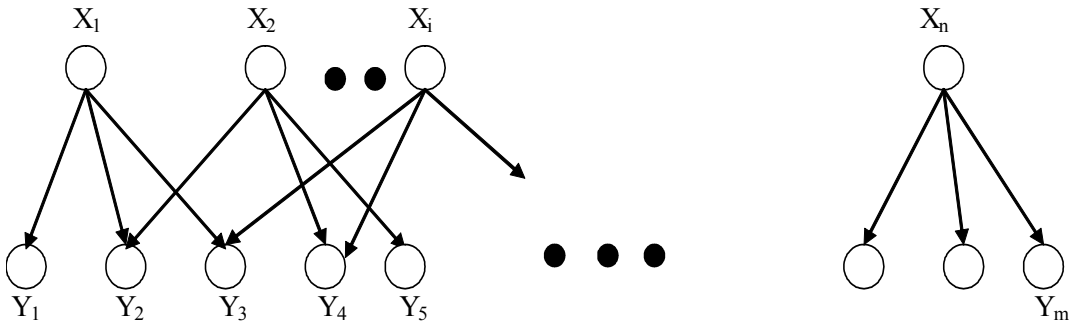


Figure 61. Bipartite graph

For the j^{th} link X_j as in Figure 61, the joint probability can be bound as

$$P(X_j = x_j^*, Y = y) = \max_{x_j} \left\{ P(X_j) \prod_{y_i \in ch_j} \left[P(y_i^+ \mid pa_{y_i^+} \right]^{y_i} \left[P(y_i^- \mid pa_{y_i^-} \right]^{1-y_i} \right\} \le \phi_f \left\{ \prod_{\substack{i=1 \\ y_i \in ch_j}}^{|ch_j|} [\eta_1]^{y_i} [\eta_0]^{1-y_i} \right\}. \quad (86)$$

As before, using the tangent plane as a bound, and the conjugate function for a NOISY-OR network, we have

$$P(X_j = x_j^*, Y = y) \le \phi_f \left\{ \prod_{\substack{i=1 \\ y_i \in ch_j}}^{|ch_j|} \left[e^{-\theta_{i0}} \right]^{y_i} \left[1 - e^{\xi_i (\theta_{i0} + r_{\max} \theta_{ij}) - f^*(\xi_i)} \right]^{1-y_i} \right\}. \quad (87)$$

Assuming homogenous conditions, (87) simplifies to

$$P(X_j = x_j^*, Y = y) \le \phi_f \left\{ \prod_{\substack{i=1 \\ y_i \in ch_j}}^{|ch_j|} \left[e^{-\theta_0} \right]^{y_i} \left[1 - e^{\alpha} \right]^{1-y_i} \right\} = \phi_f \sum_{k=0}^{|ch_j|} \binom{|ch_j|}{k} \eta_0^k \eta_1^{|ch_j|-k} = \phi(\eta_0 + \eta_1)^{|ch_j|}, \quad (88)$$

where $|ch_j|$ is the cardinality of the set of children of link *j*. Now the average *BER* can be lower bounded by the minimum *BER* over all the links *j* as follows,

$$BER = \frac{1}{n_f} \sum_{j=1}^{n_f} P(X_j \ne x_j^*(Y)) \ge \min_j BER(X_j) \ge 1 - \phi_f (\eta_0 + \eta_1)^{\max_j |ch_j|} \quad (89)$$

APPENDIX D

Proof of Corollary 2:

As in the case of MPE error, we bound the BER by a performance bound δ_{f_2}, and hence from (89), we have

$$m_f / n_f \geq \frac{\log(1-\delta_{f_2}) - \log(\phi_f)}{r_{\max}(e^{-\theta_0} + e^{\alpha})}$$
$$= \frac{\log(1-\delta_{f_2}) - \log(\phi_f)}{r_{\max} \log\left[(1-p_0) + (1-p_0)^{-\xi^{opt}} (1-p)^{-r\xi^{opt}} \xi^{opt\,\xi^{opt}} (1+\xi^{opt})^{-(1+\xi^{opt})} \right]}. \quad (90)$$

APPENDIX E

Proof of Lemma 1:

From the relation $I(\boldsymbol{X};Y) = H(Y) - H(Y \mid \boldsymbol{X})$, $H(Y \mid X) = 0$, since given the status of the links on the path of *y*, the outcome is deterministic if we assume that the network is noise free. Hence $I(\boldsymbol{X};Y) = H(\rho_{out})$, where $\rho_{out} = \overline{\rho}$. In the presence of noise we have $I(\boldsymbol{X};Y) \le H(\rho_{out})$, and hence the result can be proved.

APPENDIX F

Proof of Lemma 2:

Proof: For m_f probes, the information gained can be bounded as

$$\begin{aligned} I(\boldsymbol{X};Y_1,Y_2,.....Y_{mf}) &= H(Y_1,Y_2,.....Y_{m_f}) - H(Y_1,Y_2,.....Y_{m_f} \mid \boldsymbol{X}) \\ &\le \sum_{i=1}^{m_f} H(Y_i) = \sum_{i=1}^{m_f} H((1-(1-\rho_f)^{r_i})(1-\rho_1)+(1-\rho_f)^{r_i}\rho_2) \end{aligned} \quad (91)$$

In the absence of noise, we have $I(\boldsymbol{X};Y_1,Y_2,.....Y_{m_f}) = \sum_{i=1}^{m_f} H((1-(1-\rho_f)^{r_i}))$.

APPENDIX G

Proof of Theorem 3:

From lemma 2, for the noisy case, we have

$$I(\boldsymbol{X};\boldsymbol{Y}) \le m_f I(\boldsymbol{X};Y) \le m_f H(\rho_{out}). \tag{92}$$

The LHS of equation (92) is nothing but $n_f H(\rho_f)$, since this the number of bits required to represent the source with prior probability ρ_f [88]. Hence, we have

$$n_f H(\rho_f) \le m_f H(\rho_{out}) \Rightarrow m_f / n_f \ge H(\rho_f)/H(\rho_{out}). \tag{93}$$

In the absence of noise, the denominator can be maximized using the optimal probe route length. This is obtained by maximizing the mutual information, i.e., $\max I(\boldsymbol{X};Y) = H((1-\rho_f)^{r_{opt}}) = 1$, and hence the result follows.

APPENDIX H

Proof for optimal probe length:

Using the relation $I(\boldsymbol{X};Y) = H(Y) - H(Y \mid \boldsymbol{X})$, we have

$$\begin{aligned} H(Y \mid \boldsymbol{X}) &= -\sum_{x} P(x) \sum_{y} P(y \mid x) \log P(y \mid x) \\ &= -\sum_{x} \prod_{j=1}^{k} P(x_j) \left[P(y=1 \mid x) \log P(y=1 \mid x) + P(y=0 \mid x) \log P(y=0 \mid x) \right] \end{aligned} . \quad (94)$$

This simplifies to $H(Y \mid \boldsymbol{X}) = \bar{\rho}\left[H(\rho_2)\right] - \bar{\rho}\rho_f / (1-\rho_f)\left[H(1-\rho_1)\right]$. Also,

$$H(Y) = H\left(\bar{\rho}(1-\rho_2) + \bar{\rho}\rho_f / (1-\rho_f)\rho_1\right). \quad (95)$$

Hence, the optimum probe route length is obtained by maximizing $I(\boldsymbol{X};Y)$, and the relation can be obtained.

APPENDIX I

Proof of Proposition 1:

Part of the result is found in [86]; we provide it here for clarity. Let $\boldsymbol{X} = \left\{X_1, X_2, ...X_{n_f'}\right\}$ be the independent random variables drawn from an unknown density function $\rho_f(x)$. The joint distribution $\rho_f(\boldsymbol{X}^{n_f'}) = \prod_{i=1}^{n_f'} \rho_f(X_i)$; $n_f' = 1,2,....$. Let $\Gamma_{n_f'}$ be the space of probability distribution functions $\hat{\rho}_f(x)$. From [86], the redundancy of a code, which is the difference between the actual and ideal code lengths divided by the sample size is given by

$$\Delta_{n_f'}(\rho_f) = \frac{1}{n_f'} E\left(B(\boldsymbol{X}^{n_f'}) - \log \frac{1}{\rho_f(\boldsymbol{X}^{n_f'})} \right) \tag{96}$$

where $B(\boldsymbol{X}^{n_f'}) = \min_{\hat{\rho}_f \in \Gamma_{n_f'}} \left(L_{n_f'}(\hat{\rho}_f) + (\frac{1}{\hat{\rho}_f(\boldsymbol{X}^{n_f'})}) \right)$ is the complexity of the data $\boldsymbol{X}^{n_f'}$ relative to the code length $L_{n_f'}$ and $\Gamma_{n_f'}$. $\tilde{\rho}_{n_f'} = \arg\min_{\hat{\rho}_f \in \Gamma_{n_f'}} \left(L_{n_f'}(\hat{\rho}_f) + \frac{1}{\hat{\rho}_f(\boldsymbol{X}^{n_f'})} \right)$ is the minimum complexity estimator of the density relative to $L_{n_f'}$ and $\Gamma_{n_f'}$. Now,

$$\frac{1}{n_f'} \left(B(\boldsymbol{X}^{n_f'}) - \log(\frac{1}{\rho_f(\boldsymbol{X}^{n_f'})} \right) = \min_{\hat{\rho}_f \in \Gamma_{n_f'}} \left(\frac{1}{n_f'} L_{n_f'}(\hat{\rho}_f) + \frac{1}{n_f'} \log \frac{\rho_f(\boldsymbol{X}^{n_f'})}{\hat{\rho}_f(\boldsymbol{X}^{n_f'})} \right) \tag{97}$$

Taking the expected value w.r.t to ρ_f, we have

$$\begin{aligned} \Delta_{n_f'}(\rho_f) &= \frac{1}{n_f'} E\left(B(\boldsymbol{X}^{n_f'}) - \log(\frac{1}{\rho_f(\boldsymbol{X}^{n_f'})} \right) \\ &= E \min_{\hat{\rho}_f \in \Gamma_{n_f'}} (.) \le \min_{\hat{\rho}_f \in \Gamma_{n_f'}} E(.) = \min_{\hat{\rho}_f \in \Gamma_{n_f'}} \left(\frac{1}{n_f'} L_{n_f'}(\hat{\rho}_f) + D(\rho_f \,\|\, \hat{\rho}_f) \right) \end{aligned} . \tag{98}$$

Hence $\Delta_{n_f'}(\rho_f) \le \Omega_{n_f'}(\rho_f)$, where $\min_{\hat{\rho}_f \in \Gamma_{n_f'}} E(.) = \Omega_{n_f'}(\rho_f)$ is the penalty term called the index of resolvability in [86]. If we do not know ρ_f, then asymptotically, $n_f' \left(H(\rho_f) + \Omega_{n_c'} \right)$ bits would be needed to describe $\boldsymbol{X}$ on an average. Combining the above result from equation (98) with (93), we have the result.

APPENDIX J

Consider two samples $S_{\hat{Y}_i}$ and $S_{\hat{N}_i}$ with the range R_S. Our goal is to estimate the probability density functions from the pmf's $s_Y(k)$ and $s_N(k)$, defined over the same set of bins T, where $s_Y(k)$ and $s_N(k)$ are the samples from the k^{th} bin of $S_{\hat{Y}_i}$ and $S_{\hat{N}_i}$, respectively. Also, $\sum_{k\in T} s_Y(k) = 1$ and $\sum_{k\in T} s_N(k) = 1$. The Freedman-Diaconis rule gives the number of bins as $w = 2n^{-1/3}IQ$, where w is the bin width, n is the sample size of $S_{\hat{Y}_i} \cup S_{\hat{N}_i}$, and IQ is the interquartile range of R_S [150]. One of the problems in choosing the bins using such a rule is that the some of the bins may not include any sample, and hence the likelihood ratio, $\log_2 \frac{s_Y(k)}{s_N(k)}$, cannot be accurately estimated. The heuristic solution is to merge the bin with less ($<1\%$ of the sample size) or no samples to the adjacent bin to its right and repeat until the constraint ($>1\%$ of the sample size) is met.

REFERENCES

[1] Y. Breitbart, C.-Y. Chan, M. Garofalakis, R. Rastogi, and A. Silberschatz, "Efficiently Monitoring Bandwidth and Latency in IP Networks," *Proceedings of IEEE INFOCOM*, 2000.

[2] D. Katabi, I. Bazzi, and X. Yang, "A Passive Approach for Detecting Shared Bottlenecks," *IEEE International Conference on Computer Communications and Networks, ICCCN 2001, Arizona*, 2001.

[3] A. Markopoulou, G. Iannaccone, S. Bhattacharyya, C.-N. Chuah, and C. Diot, "Characterization of Failures in an IP Backbone," *IEEE INFOCOM*, 2004.

[4] D. Rubenstein, J. Kurose, and D. Towsley, "Detecting Shared Congestion of Flows Via End-to-end Measurement," *Proceedings of ACM SIGMETRICS'00*, June 2000.

[5] C. Ji and A. Elwalid, "Measurement-Based Network Monitoring: Achievable Performance and Scalability," *IEEE Journal of Selected Areas of Communication:Special Issue on Recent Advances in Fundamentals of Network Management*, vol. 20, pp. 714-725, May 2002.

[6] http://www.caida.org/tools/measurement/skitter/ (10/2007)

[7] http://www.caida.org/tools/measurement/skitter/references/skitter-usage-caida.xml (10/2007)

[8] D. G. Andersen, H. Balakrishnan, M. F. Kaashoek, and R. Morris, "Resilient Overlay Networks," *Proceedings of the 18th ACM SOSP, Banff, Canada*, October 2001.

[9] J. R. McIntosh, *Cellular Electron Microscopy*: Academic Press, (Editor) 2007.

[10] http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/snmp.htm (10/2007)

[11] www.cisco.com/warp/public/732/Tech/nmp/netflow/index.shtml (10/2007)

[12] http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/rmon.htm (10/2007)

[13] A. Habib, M. Khan, and B. Bhargava, "Edge-to-edge measurement-based distributed network monitoring," *Computer Networks*, vol. 44, pp. 211-233, Feb 2004.

[14] S. R. Kulkarni, G. Lugosi, and S. S. Venkatesh, "Learning Pattern Classification-A Survey," *IEEE Transactions on Information Theory*, vol. 44, pp. 2178-2206, October 1998.

[15] M. Steinder and A. S. Sethi, "Probabilistic fault localization in communication systems using belief networks," *IEEE Transactions on Networking*, vol. 5, pp. 809-822, October 2004.

[16] I. Rish, M. Brodie, S. Ma, N. Odintsova, A. Beygelzimer, G. Grabarnik, and K. Hernandez, "Adaptive diagnosis in distributed systems," *IEEE Transactions on Neural Networks*, vol. 16, pp. 1088-1109, September 2005.

[17] Y. Mao, F. R. Kschischang, B. Li, and S. Pasupathy, "A Factor Graph Approach to Link Loss Monitoring in Wireless Sensor Networks," *IEEE JSAC*, vol. 23, pp. 820-829, April 2005.

[18] R. Castro, M. Coates, G. Liang, R. Nowak, and B. Yu, " Internet Tomography: Recent Developments," *Statistical Science*, 2003.

[19] Y. Vardi, "Network tomography: Estimating source-destination traffic intensities from link data," *Journal of the American Statistical Association*, vol. 91, pp. 365–377, 1996.

[20] Y. Chen, D. Bindel, H. Song, and R. H. Katz, "An Algebraic Approach to Practical and Scalable Overlay Network Monitoring," *ACM SIGCOMM*, 2004.

[21] M. I. Jordan, *Learning in Graphical Models*: The MIT Press (Adaptive Computation and Machine Learning), 1998.

[22] T. S. Jaakkola, "Variational Methods for Inference and Estimation in Graphical Models," in *Department of Brain and Cognitive Sciences*, vol. Ph.D. Cambridge: Massachusetts Institute of Technology, 1997.

[23] S. Subramaniam and J. L. S. Milne, "Three-dimensional electron microscopy at molecular resolution," *Annual Review of Biophysics and Biomolecular Structure*, pp. 141-155, 2004.

[24] A. S. Frangakis and F. Forster, "Computational exploration of structural information from cryo-electron tomograms," *Curr Opin Struct Biol*, vol. 14, pp. 325-31, 2004.

[25] R. Narasimha, I. Aganj, M. Borgnia, G. Sapiro, S. McLaughlin, J. Milne, and S. Subramaniam, "From Gigabytes to Bytes: Automated denoising and feature identification in electron tomograms of intact bacterial cells," *IEEE International Symposium on Biomedical Imaging (ISBI '07)*, 2007.

[26] H. Winkler, "3D reconstruction and processing of volumetric data in cryo-electron tomography," *J Struct Biol*, vol. 157, pp. 126-37, 2007.

[27] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York: Wiley, 1991.

[28] J. Frank, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*: Oxford University Press, USA, 2006.

[29] J. T. Thong, K. S. Sim, and J. C. Phang, "Single-image signal-to-noise ratio estimation," *Scanning*, vol. 23, pp. 328-36, 2001.

[30] http://dsd.lbl.gov/NCS/generic/net-tools.html (10/2007)

[31] C. Labovitz, G. R. Malan, and F. Jahanian, "Internet Routing Instability," *IEEE/ACM Transactions on Networking*, vol. 6, pp. 515-528, October 1998.

[32] S. Iyer, S. Bhattacharyya, N. Taft, and C. Diot, "An approach to alleviate link overload as observed on an IP backbone," *INFOCOM*, 2003.

[33] C. Boutremans, G. Iannaccone, and C. Diot, "Impact of link failures on VoIP performance," *In Proceedings of ACM NOSSDAV*, 2002.

[34] Z. Cataltepe and P. Moghe, "Characterizing Nature and Location of Congestion on the Public Internet," *IEEE Symposium on Computers and Communication, Kemer, Antalya, Turkey*, 2003.

[35] A. Tachibana, S. Ano, T. Hasegawa, M.Tsuru, and Y. Oie, "Empirical Study on Locating Congested Segments over the Internet Based on Multiple End-to-End Path Measurements," *The 2005 Symposium on Applications and the Internet*, pp. 342 - 351, 31-04 Jan. 2005.

[36] N. Hu, L. E. Li, Z. M. Mao, P. Steenkiste, and J. Wang, "Locating Internet Bottlenecks: Algorithms, Measurements, and Implications.," *SIGCOMM*, 2004.

[37] A. Akella, S. Seshan, and A. Shaikh, "An Empirical Evaluation of WideArea Internet Bottlenecks," *In Proceedings of ACM SIGCOMM Internet Measurement Conference (IMC), Miami, FL*, October 2003.

[38] S. Kandula, D. Katabi, and J.-P. Vasseur, "Shrink: A Tool for Failure Diagnosis in IP Networks," *MineNet Workshop, SIGCOMM 2005*, Aug 2005.

[39] C. Tang, P. K. McKinley, and J. Shapiro, "Collaborative Path Selection for Topology-Aware Overlay Path Monitoring in the Presence of Selfish Agents," *Technical Report MSU-CSE-04-42*, October 2004.

[40] D. Chua, E. D. Kolaczyk, and M. Crovella, "A statistical framework for efficient monitoring of end-to-end network properties," *ACM SIGMETRICS Performance Evaluation Review*, vol. 33, pp. 390 - 391, June 2005.

[41] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution and Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721-741, 1984.

[42] S. Geman and K. Kochanek, "Dynamic Programming and graphical representation of error-correcting codes," *IEEE Transactions on Information Theory*, vol. 47, pp. 549-568, 2001.

[43] S. Klinger, S. Yemini, Y. Yemini, D. Ohsie, and S. Stolfo, "A coding approach to event correlation," *Proceedings of the fourth international symposium on Integrated network management IV*, pp. 266 - 277, 1995.

[44] I. Katzela and M. Schwartz, "Schemes for Fault identification in communication network," *IEEE Transactions on Networking*, vol. 3, December 1995.

[45] Y. Wen, V. W. S. Chan, and L. Zheng, "Efficient Fault Diagnosis Algorithms for All-Optical WDM Networks," *IEEE/OSA Journal of Lightwave Technology Special Issue on Optical Networks*, October 2005.

[46] R. G. Gallager, "Low Density Parity Check Codes," *MIT Press, Cambridge, MA*, 1963.

[47] S. B. Wicker, *Error Control Systems for Digital Communication and Storage*. Upper Saddle River, New Jersey: Prentice Hall, 1995.

[48] T. Richardson and R. Urbanke, "The Capacity of Low-Density Parity Check Codes under Message-Passing Decoding," *IEEE Transactions on Information Theory*, vol. 47, pp. 599-618, Feb 2001.

[49] W. Ryan, "An Introduction to LDPC Codes:Tutorial," 2004.

[50] A. J. Koster, R. Grimm, D. Typke, R. Hegerl, A. Stoschek, J. Walz, and W. Baumeister, "Perspectives of Molecular and Cellular Electron Tomography," *Journal of Structual Biology*, vol. 120, pp. 276-308, 1997.

[51] A. S. Frangakis and F. Forster, "Computational exploration of structural information from cryo-electron tomograms.," *Current Opinion in Structural Biology*, vol. 14, pp. 325-331, 2004.

[52] P. Penczek, M. Marko, K. Buttle, and J. Frank, "Double-tilt electron tomography," *Ultramicroscopy*, vol. 60, pp. 393-410, 1995.

[53] P. Gilbert, "Iterative methods for the three-dimensional reconstruction of an object from projections," *Journal of Theoretical Biology*, vol. 36, pp. 105-17, July 1972.

[54] R. Narasimha, A. Bennett, D. Zabransky, R. Sougrat, S. W. McLaughlin, and S. Subramaniam, "Gigabytes to Bytes: Automated Denoising and Feature Extraction as Applied to the Analysis of HIV Architecture and Variability using Electron Tomography," *IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[55] R. Narasimha, I. Aganj, A. Bennett, M. Borgnia, D. Zabransky, G. Sapiro, S. W. McLaughlin, J. L. S. Milne, and S. Subramaniam, "Quantitative Evaluation of Denoising Algorithms for Biological Electron tomography," *Journal of Structural Biology, submitted*, 2007.

[56] J. R. Kremer, D. N. Mastronarde, and J. R. McIntosh, "Computer visualization of three-dimensional image data using IMOD," *J Struct Biol*, vol. 116, pp. 71-6, 1996.

[57] M. Borgnia, S. Subramaniam, and J. Milne, "Rapid shape variation in the predatory bacterium revealed by cryo-electron tomography," *(submitted)*, 2007.

[58] J. A. Heymann, M. Hayles, I. Gestmann, L. A. Giannuzzi, B. Lich, and S. Subramaniam, "Site-specific 3D imaging of cells and tissues with a dual beam microscope," *Journal of Structual Biology*, vol. 155, pp. 63-73, 2006.

[59] A. W. J. Heymann, D. Shi, S. Kim, D. Bliss, L. S. Jacqueline, and S. Subramaniam, "Nanometer Resolution 3D Imaging of Mammalian Cells With Ion-Abrasion Scanning Electron Microscopy," *submitted*, 2007.

[60] R. E. Sockett and C. Lambert, "Bdellovibrio as therapeutic agents: a predatory renaissance?" *Nature Reviews Microbiology 2*, pp. 669-675, August 2004.

[61] R. Sougrat, A. Bartesaghi, J. D. Lifson, A. E. Bennett, J. W. Bess, D. J. Zabransky, and S. Subramaniam, "Electron Tomography of the Contact between T Cells and SIV/HIV-1: Implications for Viral Entry," *PLoS Pathogones*, vol. 3, pp. 571-581, May 2007.

[62] http://micro.magnet.fsu.edu/cells/mitochondria/mitochondria.html (11/2007)

[63] K. V. D. Bossche, J.-M. Naeyaert, and J. Lambert, "The quest for the mechanism of melanin transfer," *Traffic*, vol. 7, pp. 769-78, 2006.

[64] G. R. Rosania, "Mitochondria Give Cells a Tan," *Chemistry and Biology*, vol. 12, pp. 412-413, 2005.

[65] A. Schroeder, Y. Avnir, S. Weisman, Y. Najajreh, A. Gabizon, Y. T. J. Kost, and Y. Barenholz, "Controlling Liposomal Drug Release with Low Frequency Ultrasound: Mechanism and Feasibility," *Langmuir*, vol. 23, pp. 4019 -4025, 2007.

[66] S. A. Abraham, D. N. Waterhouse, L. Mayer, P. Cullis, T. Madden, and M. B. Bally, "Liposomal Formulations of Doxorubicin," in *Methods in Enzymology, Liposomes, Part E*, vol. 391, N. Duzgunes, Ed.: Elsevier Academic Press, 2005, pp. 71-97.

[67] S. K. Hobbs, W. L. Monsky, F. Yuan, W. G. Roberts, L. Griffith, V. P. Torchilin, and R. K. Jain, "Regulation of transport pathways in tumor vessels: Role of tumor type and microenvironment," *Proceedings of The National Academy of Science, USA*, vol. 95, pp. 4607-4612, 1998.

[68] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, pp. 321-331, 1987.

[69] C. Xu and J. L. Prince, "Snakes, Shapes, and Gradient Vector Flow," *IEEE Transactions on Image Processing*, vol. 7, pp. 359-369, March 1998.

[70] V. Caselles, F. Catte, T. Coll, and F. Dibos, "A Geometric Model for Active Contours in Image Processing," *Numerische Mathematik*, vol. 66, pp. 1-31, October 1993.

[71] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *International Journal of Computer Vision*, vol. 22, pp. 61 -79, 1997.

[72] R. Malladi, J. A. Sethian, and B. C. Vemuri, "Shape Modeling with Front Propagation: A Level Set Approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 158-175, 1995.

[73] D. Mumford and J. Shah, "Boundary detection by minimizing functionals," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 22--26, 1985.

[74] A. Tsai, A. Yezzi, and A. S. Willsky, "Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation, and magnification," *IEEE Transactions on Image Processing*, vol. 10, pp. 1169 -1186, 2001.

[75] T. F. Chan and L. A. Vese, "Active Contours without Edges," *IEEE Transactions on Image Processing*, vol. 10, pp. 266-277, 2001.

[76] V. Baba and C. Yunmei, "Joint image registration and segmentation," in *Geometric level set methods in imaging, vision, and graphics*. New York: Springer, 2003, pp. 251--269.

[77] V. Paxson, "End-to-End Routing Behavior in the Internet," *IEEE/ACM Transactions on Networking*, vol. 5, pp. 601-615, October 1997.

[78] M. Brodie, I. Rish, and S. Ma, "Intelligent Probing: A cost-effective approach to fault diagnosis in computer networks," *IBM Systems Journal*, vol. 41, 2002.

[79] R. Castro, M. Coates, G. Liang, R. Nowak, and B. Yu, "Internet Tomography: Recent Development," *Statistical Science*, 2003.

[80] S. Seetharaman and M. Ammar, "On the Interaction between Dynamic Routing in the Overlay and Native Layers," *IEEE INFOCOM*, April, 2006.

[81] http://www.routeviews.org/ (8/2007)

[82] A. Medina, A. Lakhina, I. Matta, and J. Byers, "BRITE: Universal Topology Generation from a user's Perspective," *Technical Report (User Manual) BU-CS-TR-2001-003*, April 2001.

[83] V. V. Vazirani, *Approximation Algorithms*: Springer, 2004.

[84] R. Narasimha, S. Dihidar, C. Ji, and S. W. McLaughlin, "Scalable Fault Detection in IP Networks using Variational Inference Approach," *IEEE Communication Conference (ICC)*, 2007.

[85] J. Rissanen, "Universal coding, information, prediction and estimation," *IEEE Transactions on Information Theory*, vol. 30, pp. 629-636, July 1984.

[86] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Transactions on Information Theory*, vol. 37, pp. 1034-1054, 1991.

[87] V. Paxson, "End-to-End Internet Packet Dynamics," *IEEE/ACM Transactions on Networking*, vol. 7, pp. 277-292, June 1999.

[88] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, August 1991.

[89] D. Burshtein, M. Krivelevich, S. Litsyn, and G. Miller, "Upper Bounds on the Rate of LDPC Codes," *IEEE Transactions on Information Theory*, vol. 48, pp. 2437-2449, Sept 2002.

[90] R. G. Gallager, *Information Theory and Reliable Communication*: Wiley, 1968.

[91] R. Narasimha, S. Dihidar, C. Ji, and S. W. McLaughlin, "A Scalable Approach to Localize Congestion using Coding Theory," *Proceedings of Forty-third Allerton Conference on Computers, Control and Communications*, 2005.

[92] R. Narasimha, S. Dihidar, C. Ji, and S. W. McLaughlin, "A Scalable Probing-based Approach for Congestion Detection using Message Passing," *Proceedings of ICC*, 2006.

[93] S. Floyd, "TCP and Explicit Congestion Notification," *ACM Computer Communication Review*, vol. 24, pp. 10-23, October 1994.

[94] S. Floyd and V. Jacobson, "Random Early Detection gateways for Congestion Avoidance," *IEEE/ACM Transactions on Networking*, vol. 1, pp. 397-413, August 1993.

[95] R. Narasimha, S. Dihidar, C. Ji, and S. W. McLaughlin, "Scalable probing-based monitoring in IP networks using measurements and inference: A learning Framework," *IEEE Transactions on Signal Processing, revised version submitted*, 2007.

[96] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, pp. 629--639, 1990.

[97] A. S. Frangakis, A. Stoschek, and R. Hegerl, "Wavelet transform filtering and nonlinear anisotropic diffusion assessed for signal reconstruction performance on multidimensional biomedical data," *IEEE Trans Biomed Eng*, vol. 48, pp. 213-22, 2001.

[98] G. Gilboa, N. Sochen, and Y. Y. Zeevi, "Image Enhancement and Denoising by Complex Diffusion Processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 26, pp. 1020-36, 2004.

[99] D. L. Donoho, "Denoising via soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, pp. 613-627, 1995.

[100] P. Kovesi, "Phase Preserving Denoising of Images," *The Australian Pattern Recognition Society Conference: DICTA'99*, pp. 212-217, 1999.

[101] J. Portilla, V. Strela, M. Wainwright, and E. P. Simoncelli, "Image Denoising using Scale Mixtures of Gaussians in the Wavelet Domain," *IEEE Trans Image Process*, vol. 12, pp. 1338-1351, 2003.

[102] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Boston, Massachusetts: Addison-Wesley, 1992.

[103] J. J. Fernandez and S. Li, "An improved algorithm for anisotropic nonlinear diffusion for denoising cryo-tomograms," *J Struct Biol*, vol. 144, pp. 152-61, 2003.

[104] W. C. Moss, S. Haase, J. M. Lyle, D. A. Agard, and J. W. Sedat, "A novel 3D wavelet-based filter for visualizing features in noisy biological data," *J Microsc*, vol. 219, pp. 43-9, 2005.

[105] R. S. Pantelic, R. Rothnagel, C. Y. Huang, D. Muller, D. Woolford, M. J. Landsberg, A. McDowall, B. Pailthorpe, P. R. Young, J. Banks, B. Hankamer, and G. Ericksson, "The discriminative bilateral filter: an enhanced denoising filter for electron microscopy data," *J Struct Biol*, vol. 155, pp. 395-408, 2006.

[106] W. Jiang, M. L. Baker, Q. Wu, C. Bajaj, and W. Chiu, "Applications of a bilateral denoising filter in biological electron microscopy," *J Struct Biol*, vol. 144, pp. 114-22, 2003.

[107] J. Weickert, *Anisotropic Diffusion in Image Processing*. Stuttgart, Germany: Teubner-Verlag, 1998.

[108] http://www-stat.stanford.edu/~wavelab/WaveLab701.html (8/2006)

[109] http://decsai.ugr.es/~javier/denoise/software/index.htm (8/2006)

[110] A. S. Frangakis and R. Hegerl, "Noise reduction in electron tomographic reconstructions using nonlinear anisotropic diffusion," *J Struct Biol*, vol. 135, pp. 239-50, 2001.

[111] K. Sandberg, "Methods for image segmentation in cellular tomography," *Methods Cell Biol*, vol. 79, pp. 769-98, 2007.

[112] A. S. Frangakis and R. Hegerl, "Segmentation of two- and three-dimensional data from electron microscopy using eigenvector analysis," *J Struct Biol*, vol. 138, pp. 105-13, 2002.

[113] K. Sandberg and M. Brega, "Segmentation of thin structures in electron micrographs using orientation fields," *J Struct Biol*, vol. 157, pp. 403-15, 2007.

[114] M. Jiang, Q. Ji, and B. F. McEwen, "Automated extraction of fine features of kinetochore microtubules and plus-ends from electron tomography volume," *IEEE Trans Image Process*, vol. 15, pp. 2035-48, 2006.

[115] http://www.tgs.com/ (8/2006)

[116] P. Gilbert, "Iterative methods for the three-dimensional reconstruction of an object from projections," *J Theor Biol*, vol. 36, pp. 105-17, 1972.

[117] J. Winn, A. Criminisi, and T. Minka, "Object Categorization by Learned Universal Visual Dictionary," *Proceedings of the 2005 IEEE International Conference on Computer Vision (ICCV), Beijing*, 2005.

[118] G. Csurka, C. Bray, C. R. Dance, and L. Fan, "Visual Categorization with Bags of Keypoints," *Proceedings of the 8th European Conference on Computer Vision (ECCV), Prague*, May 2004.

[119] M. Varma and A. Zisserman, "A Statistical Approach to Texture Classification from Single Images," *International Journal of Computer Vision*, vol. 62, pp. 61-81, April 2005.

[120] T. Leung and J. Malik, "Recognizing Surfaces Using Three-Dimensional Textons," *Proceedings of the IEEE Conference on Computer Vision (ICCV), Greece*, 1999.

[121] O. G. Cula and K. J. Dana, "Compact representation of bidirectional texture functions," *Proceedings of Computer Vision and Pattern Recognition*, vol. 1, pp. 1041-1047, 2001.

[122] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study," *International Journal of Computer Vision*, vol. 73, 2007.

[123] K. J. Dana, B. v. Ginneken, S. K. Nayar, and J. J. Koenderink, "Reflectance and Texture of Real-World Surfaces," *ACM Transactions on Graphics*, vol. 18, pp. 1-34, 1999.

[124] S. Lazebnik, C. Schmid, and J. Ponce, "A Sparse Texture Representation Using Local Affine Regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1265-1278, 2004.

[125] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models for 101 object categories," *Computer Vision and Image Understanding*, 2007.

[126] R. Fergus, P. Perona, and A. Zisserman, "Object Class Recognition by Unsupervised Scale-Invariant Learning," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 264-271, 2003.

[127] M. R. Everingham, A. Zisserman, C. K. I. Williams, and L. V. Gool, "The 2005 PASCAL Visual Object Classes challenge," in *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, J. Quinonero-Candela, I. Dagan, B. Magnini, and F. d'Alche-Buc, Eds.: Springer-Verlag, LNAI 3944, 2006, pp. 117-176.

[128] J. Kaufhold, R. Collins, A. Hoogs, and P. Rondot, "Recognition and Segmentation of Scene Content using Region-Based Classification," *Proceedings of the 18th International Conference on Pattern Recognition*, vol. 1, pp. 755 - 760, 2006.

[129] T. Leung and J. Malik, "Representing and Recognizing the Visual Appearance of Materials using Three-dimensional Textons," *International Journal of Computer Vision*, vol. 43, pp. 29 - 44, June 2001.

[130] O. G. Cula and K. J. Dana, "3D Texture Recognition Using Bidirectional Feature Histograms," *International Journal of Computer Vision*, vol. 59, pp. 33 - 60, August 2004.

[131] C. Schmid, "Constructing Models for Content-Based Image Retrieval," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2001.

[132] M. Varma and A. Zisserman, "Texture Classification: Are Filter Banks Necessary?" *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 691-697, 2003.

[133] A. A. Efros and T. K. Leung, "Texture Synthesis by Non-parametric Sampling," *IEEE International Conference on Computer Vision, Corfu, Greece*, pp. 1033-1038, 1999.

[134] U. Grenander and M. I. Miller, "Representation of knowledge in complex systems," *Journal of Royal Statistical Society (B)*, vol. 56, pp. 1-33, 1994.

[135] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Appearance, Shape and Context," *Invited submission to appear in International Journal on Computer Vision (IJCV), special issue*, 2007.

[136] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to Detect Natural Image Boundaries Using Brightness, Color and Texture Cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 1-20, 2002.

[137] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and Texture Analysis for Image Segmentation," *International Journal of Computer Vision*, vol. 43, pp. 7-27, 2001.

[138] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition," *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, 2006.

[139] O. Tuzel, L. Yang, P. Meer, and D. J. Foran, "Classification of Hematologic Malignancies using Texton Signatures," *Pattern Analysis & Applications*, 2006.

[140] E. N. Issam, Y. Y. Yang, N. Miles, P. Galatsanos, and R. M. Nishikawa, "A support vector machine approach for detection of microcalcifications," *IEEE Transactions on Medical Imaging*, vol. 21, pp. 1552-1563, 2002.

[141] D. A. Adjeroh, U. Kandaswamy, and J. V. Odom, "Texton-based segmentation of retinal vessels," *Journal of Optical Society of America (A)*, vol. 24, pp. 1384-1393, May 2007.

[142] G. Gilboa, N. Sochen, and Y. Y. Zeevi, "Variational denoising of partly-textured images by spatially varying constraints," *IEEE Transactions on Image Processing*, vol. 15, pp. 2281-2289, 2006.

[143] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, pp. 259-268, November 1992.

[144] S. Osher and N. Paragios, "Geometric Level Set Methods in Imaging, Vision, and Graphics," Springer, 2003.

[145] A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, E. Grimson, and A. Willsky, "A Shape-Based Approach to Curve Evolution for Segmentation of Medical Imagery," *IEEE Transactions on Medical Imaging*, vol. 22, pp. 137-154, February 2003.

[146] M. Bertalmío, G. Sapiro, and G. Randall, "Region tracking on level-Sets methods," *IEEE Transactions on Medical Imaging*, vol. 18, pp. 448-451, 1999.

[147] C. Li, C. Xu, C. Gui, and M. D. Fox, "Level Set Evolution without Re-Initialization: A New Variational Formulation," *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005.

[148] R. Narasimha, H. Ouyang, A. Gray, S. W. McLaughlin, and S. Subramaniam, "Automatic Mining of Whole Cell tomogra ms for Cancer Detection," *NIPS workshop on New Problems and Methods in Computational Biology*, 2007.

[149] B. Davie, T. Li, E. Rosen, and Y. Rekhter, "Explicit Route Support in MPLS," *draft-davie-mpls-explicit-routes-00.txt*, November 1997.

[150] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley-Interscience, 2002.

VITA

Rajesh Narasimha was born in Mumbai and was brought up in Mangalore, India. He got his Bachelor's degree in Electronics and Communication engineering from Mangalore University in June 1999. From August 1999-2000, he was a research associate at the ECE dept., Indian Institute of Science (IISc), Bangalore working on speech recognition for Ericsson Inc. In August 2002, he received a Master's degree, in electrical engineering, from the Rochester Institute of Technology, Rochester, NY under the guidance of Dr. Raghuveer Rao. From August 2002-December 2002 he was a research associate at RIT, and worked with Xerox Corporation, Webster, Rochester, NY on content-based video summarization and browsing applications. Since spring of 2003, he has been a Ph.D. student in the telecom group at Georgia Institute of Technology. From May 2006-August 2007, he was pre-doctoral fellow at the National Cancer Research Institute, NIH, Bethesda, MD, where he worked on applications of machine learning to biology. He is the recipient of SPIE scholarship for 2003-2004, and the NIH pre-doctoral fellowship for 2006-2007. He is a student member of IEEE and SPIE. He completed the requirements for the degree of Doctor of Philosophy (Ph.D.) at the Georgia Institute of Technology in November 2007. His research interests span broad areas of signal, image and video processing, machine learning applications to scalable data mining from large and complex datasets, and statistical modeling and data analysis.